万卷方法 | 社会测量与评估方法译丛

EVALUATION
A SYSTEMATIC
APROACH

第7版

评估：方法与技术

彼得·罗希
马克·李普希
霍华德·弗里曼
著

邱泽奇　王旭辉　刘　月　等
译



重庆大学出版社
http://www.cqup.com.cn

责任编辑 雷少波 林 萍
封面设计 黄河封面工作室

本书是第一本被翻译成中文的评估教材。她诞生于1970年代的美国，经过30余年的实践和修订，至今已经出版至第7版。

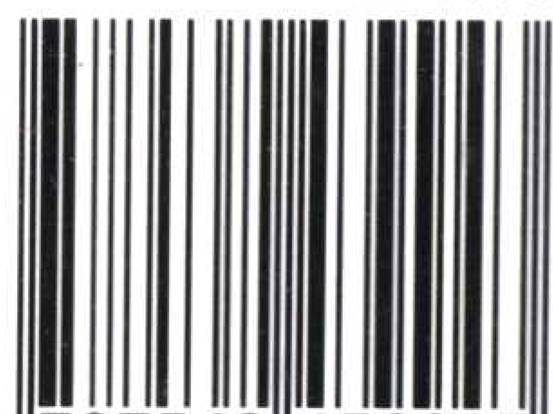
评估可以说明哪些项目更可能达到政策目标，同时也能说明哪些项目对社会和人民而言成本更小。

评估工作者来自几乎所有的社会科学领域：经济学、心理学、社会学、政治学、人类学以及教育学。评估所使用的方法和理论也来自社会科学的各个领域。

本书的核心主题是为对社会项目的设计、实施和利用进行评估的各种研究活动提供一个引导。我们试图把握评估研究的全局，分析社会项目评估的设计、实施、绩效和效率。从而为准备从事评估职业或者需要了解评估活动的人们，提供有关评估研究的基本知识和技巧，并分享在长期评估实践中所积累起来的集体经验。

本书的读者对象包括：社会科学各专业的学生和研究人员、社会项目负责人、社会评论者。

ISBN 978-7-5624-3994-3



9 787562 439943 >

定价：49.00元

万卷方法 | 社会测量与评估方法译丛

EVALUATION
A SYSTEMATIC
APROACH

第7版

评估：方法与技术

彼得·罗希
马克·李普希
霍华德·弗里曼
著
邱泽奇 王旭辉 刘 月 等
译

重庆大学出版社

版贸渝核字(2006)第 37 号

图书在版编目(CIP)数据

评估:方法与技术/(美)罗希(Rossi, P. H.),(美)李普希(Lipsey, M. W.),(美)弗里曼(Freeman, H. E.)著;邱泽奇等译. —重庆:重庆大学出版社,2007.4
(万卷方法. 社会评估与测量方法译丛)
书名原文:Evaluation: A Systematic Approach
ISBN 978-7-5624-3994-3

Ⅰ. 评… Ⅱ. ①罗…②李…③弗…④邱… Ⅲ. 评估—方法—教材 Ⅳ. C93-03

中国版本图书馆 CIP 数据核字(2007)第 027060 号

评估:方法与技术(第 7 版)
彼得·罗希 马克·李普希 霍华德·弗里曼 著
邱泽奇 王旭辉 刘 月 等译
责任编辑:雷少波 林 萍 版式设计:雷少波
责任校对:邹 忌 责任印制:张 策
*
重庆大学出版社出版发行
出版人:张鸽盛
社址:重庆市沙坪坝正街 174 号重庆大学(A 区)内
邮编:400030
电话:(023) 65102378 65105781
传真:(023) 65103686 65105565
网址:http://www.cqup.com.cn
邮箱:fxk@cqup.com.cn (市场营销部)
全国新华书店经销
重庆科情印务有限公司印刷
*
开本:787×1092 1/16 印张:19.5 字数:403千 插页:16 开 2 页
2007 年 4 月第 1 版 2007 年 4 月第 1 次印刷
印数:1—3 000
ISBN 978-7-5624-3994-3 定价:49.00 元

中文版前言

我非常荣幸地知道，我们这本书是第一本被翻译成中文的评估教材。中国是一个伟大且强大的国家，正如我们在书中解释过的一样，我希望能够借助评估研究的思路和方法促进中国社会更快地发展。

评估不能代替政治过程。政策是由政治官员和立法人员制定的。但是，政策的实施却有很多方式。评估的任务就是了解项目的形成，即形成政策的特定形式。系统地评估，可以说明哪些项目更可能达到政策目标，同时也说明哪些项目对社会、人民和社会制度而言成本更小。

评估也被应用在社会科学中。美国的评估工作者来自几乎所有的社会科学领域——经济学、心理学、社会学、政治学以及人类学。评估所使用的方法和理论也来自社会科学的各个领域。除了学科背景的差异以外，所有的评估工作者都有着共同的目标——通过自己的职业实践来改善社会环境。

我们期待着通过评估工作者的努力，中国的评估实践能够为本书未来的美国版本增添最好的评估实践案例。

罗希
李普希

作者前言

本书第7版吸纳了一些新的研究材料,并对前一版中的某些内容进行了大幅修订。本版的修正涉及:广泛地调整了产出测量(outcome measurement)和督导(monitoring)的内容;更完善地说明了项目影响评估的设计和筹划;更加全面地阐述了评估研究中的重要统计问题;此外,还对事后分析作出了更详尽的描述。我们相信,这些调整将会使本书的相关内容更贴近评估研究领域的前沿。

然而,本书的目标和核心主题始终没变——那就是为对社会项目的设计、实施和利用进行评估的各种研究活动提供一个引导。我们试图把握评估研究的全局,分析社会项目评估的设计、实施、绩效和效率。尽管做出了很多修订,但是本书的目标仍然是为准备从事评估职业或者需要了解评估领域的人们,提供有关评估研究的基本知识和技巧,并分享在长期评估实践中所积累起来的集体经验。本书的读者对象包括:学生、实际工作者、社会项目负责人、社会评论者,以及任何其他成员,只要他对意在改善社会状况的社会项目的成败测量感兴趣。

我们相信,通过这本书的阅读,您将可以获得一些必要的知识以理解和评判评估研究。应该说明的是,尽管我们对评估研究程序进行了一定的描述,并对相应的细节进行了说明,甚至提及了先前研究者们已有的研究成果和讨论;不过,本书并不想成为一本仅仅介绍评估步骤的教材。根本上讲,具体项目设计和操作中的研究经验,是这本书所不能带给广大读者的。我们鼓励所有想进入评估研究领域的读者,通过亲自实践来获取工作经验。

20世纪70年代,当本书第1版出版发行时,评估还未成为评定社会项目的一种完善方法。现在的情况则大不相同。21世纪,评估研究在全世界范围内获得了发展,评估的研究程序已经牢固地进入了世界范围内每个政府层次的日常活动中,也进入了非政府组织的运作中,甚至进入了社会议题的讨论中,成为不可缺少的一个组成部分。媒体几乎从未停止过向公众发布某些社会项目评估的结果。我们相信,评估研究对于社会政策的制定与改善,会起到十分重要的作用。成为一名评估者,将会担负起一个富有刺激性的职业角色,有机会通过运用专业技能和人际关系技巧,参与社会状况的改善。

我们将本书的第7版献给丹尼尔·帕特里克·莫怡尼安(Daniel Patrick Moynihan)。他已于最近逝世。半个世纪以来,他有着令人惊奇的个人经历,既保持着在学术领域的核心位置(哈佛大学),也在联邦政府机构中(肯尼迪和约翰逊总统执政时,他担任劳工部长助理)身居要职,还在尼克松政府中担任城市问题方面的白宫顾问,并两次当选纽约市的联邦参议员。同时,他还出版了数本有关联邦政府社会政策

和决策的有影响力的著作。莫怡尼安在参议院的工作,明显提高了参议院商议社会政策的理性程度。尽管经历丰富,但是在他从事的每份工作中,改善社会政策都是其关心的核心问题。另外,他在社会政策尤其是评估研究领域,是一名出色而坚定的鼓吹者(我们可以在第1章的专栏1—A中看到一个例子)。无论是在直接推进或者间接支持联邦政府的评估活动方面,还是在推进改善我们社会生活状况的其他项目中,他都扮演了十分关键和积极的角色。

彼得·罗希(P. H. R.)

马克·李普希(M. W. L.)

目 录

项目、政策和评估 1

本章介绍项目评估。笼统地讲,评估就意味着确定某些预期目标的价值或者将特定价值赋予到某些目标之上。在本书中,我们将评估这一概念限定在项目评估或者评估研究的范畴,项目评估是一种社会科学活动,涉及搜集、分析、解释和沟通有关旨在改善社会环境的社会项目的实施和绩效。评估有各种实用目的:帮助决定项目是否应该继续、改善、扩展或缩减;评估新项目的用途和创意;提高项目管理和指导的绩效;满足项目各方的要求。评估也有助于丰富实际的、方法论性的社会科学知识。

要理解现代背景下的评估活动,需要懂得一些评估的历史、评估特有的概念和目的、评估关注的问题和操作中所面临的内在张力与挑战。项目评估意味着把社会研究方法运用于分析一定政治和组织环境下的社会干预,而合理的评估将有助于对社会问题的评判,使项目的设计、实施、影响和绩效满足干预的需要。单个的评估研究以及许多类似研究的知识积累,能够为旨在改善人类环境的社会行动提供实质性的帮助。

自古以来,许多有组织的努力都致力于描述、理解和改善人类环境的缺陷。这本书承袭了对社会问题进行科学研究的传统——一种渴望改善我们的物质和社会环境质量、通过系统地创造和运用知识,提高个体和集体福利的传统。尽管**项目评估**(Program evaluation)和**评估研究**(Evaluation research)是新近创造的词汇,但却不是刚刚才有的活动。评估活动可以追溯到科学的初创期。科隆巴赫及其同事(Cronbach,1980)指出,三个世纪以前,霍布斯(Thomas Hobbes)及其同时代追随者曾努力用数量方法评估社会环境,探寻死亡率、发病率和社会解体的原因。

即使是社会实验,作为当代评估研究面临的最大技术挑战,也不是最近的发明。最早的"社会实验"发生在18世纪初期。那时,一位英国船长观察到,在地中海地区的诸多国家船只上工作的海员很少患坏血症;他同时注意到,柑桔类果实是这些海员的口粮之一。因此,他把自己的船员分为两半,一半人食用柠檬,另一半人则沿袭原有的饮食。实验表明,食用柠檬有助于防止坏血病。这位船长也许不知道他在评估一项示范性项目,当然,也不可能获得细致的"项目理论"(后面我们将讨论这个术语)。按照现在的医学知识可知,患坏血症是因为缺乏维生素C,而柠檬富含维生素C。总之,干预发生了作用。由此,英国海员都被迫食用柑桔类果实,这就是柠檬茶流行的由来。巧合的是,人们花了大约50年的时间才使船长的"社会项目"得到广泛利用。不过,直到现在,传播和接受评估的发现,仍然不是一件易事。

什么是评估研究

在不同时期,为了获得某些期望的结果,政策制定者、资助机构、规划者、项目管理者、纳税人或项目对象,都需要区分值得实施的和不值得实施的**社会项目**(Social program)①,启动新项目和改善既有项目,从而达到特定的预期目标。为了做到这一点,他们必须获得下面一些问题的答案:

- 问题的特质和范围是什么?问题出在哪里?影响到谁?影响了多少人?如何影响?
- 什么样的问题或结果能够说明新的、扩展的或修订的社会项目的必要性?
- 可行的、能对问题产生明显改善作用的干预是什么?
- 干预的对象是什么?
- 特定干预是否落实到了目标群体?
- 干预活动实施得好吗?提供了原定的服务吗?
- 干预对于实现预期目标或利益是否有效?
- 与绩效和收益比较,项目的成本是否恰当?

① 这是本书的一个核心概念,已经在本章的基本概念一栏中列出。

即使是局部性的、具体的项目,获得上述问题的答案也是必要的。譬如小镇的职业培训、小学新数学教学计划、社区精神健康出诊。对全国性的、全州性的项目如健康、社会福利以及教育改革而言,也是必要的。提供上述问题的答案,是项目评估工作者的责任。具体地说,评估研究者(评估者)运用社会研究方法,研究、评价,并帮助改善社会项目的所有重要方面,包括社会问题诊断、概念化与设计、实施与管理、产出,及其效率(专栏 1—A 表现了一位活跃的参议员对项目绩效评估证据的重视)。

专栏 1—A 政策制定老手要看评估结果

尽管不断有行政官员说服我们,要增加这个或那个社会项目的经费,但我们仍然在尽最大的努力(可以说,现在取得了很大的成功)[减少赤字]……在这些琐碎中,我印象最深的是"家庭维持"项目。家庭维持项目是另一类援助项目(已经有了许多了),属于社会服务的一块,并由某个附属委员会管理。该项目为期 5 年,耗资 9.3 亿美元,1994 财政年度的启动经费为 6 000 万美元。在过去的 30 年里,我看到了家庭与社会的分离;现在,每个新小组的成员都告知我,只要再加一个项目……让我冒昧地把我 1993 年 7 月 28 日写给泰森(Laura D' Andrea Tyson)博士(后任经济咨询委员会主席)的信作为有关"家庭维持"项目的档案:

亲爱的泰森博士:

您也许记得,上星期四在您参加民主党政策委员会会议时,我和您谈到了总统的家庭维持项目。您指出,他非常支持这个项目。我向您保证,我也支持这个项目,只是我希望看到一些证据,来说明项目会产生效果。对此,您明确表示已经有资料可以证明,但出于好奇,我曾要求两个佐证。

次日,我便收到了您的工作人员格赖德(Sharon Glied)发来的传真,提供了一些佐证和一篇关于"结果评估"文章,文章似乎是华盛顿社会政策研究中心的法偌(Frank Farrow)和芝加哥大学霍尔(Chapin Hall)中心的理查曼(Harold Richman)写的。文章非常直率地写道:"就家庭维持服务而言,尚缺乏能影响整个国家经费安排的实在证据。"

也就是昨天,霍尔中心又发布了"伊利诺斯家庭优先防护项目评估:总报告"。这是自 1987 年伊利诺斯家庭维持法案后,对家庭优先安置项目的一项大型研究。研究"旨在考察家庭优先安置项目的效果和其他影响,譬如随后的儿童虐待"。家庭优先安置项目的工作人员提供了大约 4 500 个案例资料和服务特点资料:大约 1 600 个家庭参与了项目的随机实验。研究结果非常清楚。

总体上,家庭优先安置项目略微提高了一点安置率(当综合所有实验资料时)。但是,如果考虑到具体案例和区域因素,这个效果就消失了。换句话说,这一项目既没有负面效果,也没有正面效果。

这个结果并不新鲜,因为在 1992 年,罗希(Peter Rossi)在"评估家庭维持项目"的结论中已经说过了。今天的评估"并没有提供充分的证据以说明家庭维持计划是否有效果"。我是否可以对您说,这些发现并没有令人惊奇的地方?从 20 世纪 60 年代中期起,这样的评估一再重复,甚至持续不断。但很少具体说明项目有正面效果或负

面效果。而20世纪70年代的负债税实验,似乎还增加了家庭解体的比例。

这种“与印象相反”的发现,最早出现在20世纪60年代。格里雷和罗希(Greeley and Rossi)以及我的部分工作,还有科尔曼(J. S. Coleman)的工作,都说明了这一点。直到今天,我仍然不能确定,我们面对的只是方法的构造,还是规模更大的和更棘手的社会项目现实。任何一种情况都印证了罗希1978年提出的铁律,即“如果说过去几十年评估活动有什么经验规律的话,那就是,社会项目测量效果的期望值等于零”。

我之所以给您写这封长信,是因为我认为,有一件事情非常重要。在过去的6个月里,克林顿政府的不少人不断试图说服我,让我知道一些社会政策已经获得了极大的效果,但是在我看来,那些社会政策没有任何效果。我认为这样的事很危险。因为政策的效果并不稳定,甚至某些人的看法从根本上错误。因此,要确切地知道什么样的项目的确没有效果。意识上的自信很容易演变为固执,进而忽视一切事实。

这个时候(甚至这一代),政治保守派的最大优势在于他们对复杂思想的开放态度。而自由派的弹性则在于抵制复杂思想。在过去的12年里,我曾经极力去改变;现在看来,这些状况不仅没有被改变,反而得到了强化。如果是这样的话,自由主义的复活将只会昙花一现,并不会产生任何结果。

丹尼尔·帕特里克·莫伊尼安 (Daniel Patrick Moynihan)参议员 敬上

资料来源:D. P. Moynihan, *Miles to Go: A Personal History of Social Policy* (Cambridge, MA: Harvard University Press, 1996), pp. 47-49.

尽管本书强调的是对社会项目(特别是人类服务项目)的评估,但是项目评估并不局限于此。美国审计总署(GAO)项目评估方法部的工作,就很好地说明了项目评估的广泛性。在这个部门的发展史上,它评估了军用设备的获得与测试、饮用水的质量控制、重要高速公路的维护、使用激素促进菜牛生长,以及其他各种有组织的和非服务性的活动。

的确,本书提供的技术基本上适用于各种有组织的社会活动所提出的绩效问题。举例而言,评估手段已经被运用于大众传媒和广告业的发展与市场开发中。商业和工业企业对选择、培训、晋升职员和组织劳动力的程序需要评估;政治人物则通过评估选民诉求来发展竞选策略;人们要对消费品性能、耐用性和安全性进行测试;公共和私有部门的管理者对职员、财务和组织的人事实践也要进行评价。当然,还有很多类似例子可以说明评估手段的广泛运用。

五花八门的评估之间的区别,就在于各种努力的特质和目标。本书强调的是评估各种旨在改善人类环境的项目,而不是评估那些旨在增加利润或扩大影响和权力的项目。之所以这样,是考虑把评估的活跃领域和实践的需要结合起来,进而限制本书的篇幅。这一点可以从我们所使用的概念中得到印证,比如**评估**、**项目评估**和**评估研究**这三个概念,它们是可以互换的。

为使项目评估更贴近现实和实际,下面提供了一些社会项目评估的案例,其中评估的主办方既有地方、州、联邦政府机构、国际组织、私营基金会和慈善机

构，也有其他非营利机构、营利机构和企业。

- 在美国一些主要城市，某一私营基金会提供了一笔启动费，用于在低收入社区建立社区医疗中心。目的是为当地居民提供流动医疗服务，以免他们花大笔的公共支出让医生出诊或进急救室。进一步的目的是，让社区居民获得就医机会，这样，既减少了就医时间，也降低了就医费用。评估结果表明，与医院门诊比较，有些医疗中心的确在成本—绩效方面具有优势。
- 在纽约市，部分对学校提供经费资助的人创立了一个私人资助项目，用于帮助在较差公立学校就读 1—3 年级的穷人孩子。符合条件的家庭可以获得奖学金资助，用于支付在任何一所私立学校就读 3 年的学费。在 14 000份申请中，项目方随机选择 1 500 人进行资助。评估小组利用了这一选择模型，认为这是一项随机实验，并试图比较获得奖学金进入私立学校念书的结果与没有获得奖学金仍然在公立学校念书的结果。
- 在过去 10 年中，联邦政府允许州政府修改福利项目，而由此对项目对象产生的影响和需要的成本，则需要评估。一些州提高了职业训练的要求，另一些州要求在一定时间内见到效果，少数州则禁止为新生婴儿提高福利。评估结果表明，这些政策可以降低福利费用和提高就业率。为此，联邦政府 1996 年通过的福利改革法案（个人责任和工作机会协调法案）吸纳了项目的许多特征。
- 全世界足有 2/3 的农村儿童营养不良或严重营养不良，这对他们的健康和身心成长构成了严重影响。中非某些地区通过提供补充营养，展现了改善儿童身心健康的巨大潜力。项目方每天向孕妇、处于哺乳期的母亲、12 岁以前的儿童提供高蛋白、高热量的食物补充。尽管这样的营养补充对身体成长的效果非常明显，但在对心智提高的影响上却表现一般。
- 为了提高工人的满意程度和生产质量，一家大型制造公司对职工进行重组，成立了一些独立的工作小组。在小组内部，工人自己指派任务，向管理部门提供产量建议，根据产量和质量改良情况，投票决定奖金的发放。项目评估所得到的资料显示：这种的做法减少了旷工率、人员调整率，并因此提高了劳动生产率。

这些短小的例子说明，系统评估工作涉及多样化的社会干预项目。但是，以上所有这些都涉及一项特别的评估活动：项目产出评估。正如后面将要讨论的，评估也可以专注于项目的需要、设计、操作、服务或者效率。

评估研究简史

尽管评估的历史可以追溯至 17 世纪，但是，系统的评估研究则是现代社会的产物，评估研究的发展是 20 世纪以来的事情。项目评估所使用的社会研究方

法,与方法本身的发展和改进,以及意识形态、政治和民主的变迁是相伴随的。

作为社会科学活动的评估研究

系统的社会项目评估最早出现在教育和公共健康领域。在第一次世界大战以前,最有意义的努力就是扫盲、职业培训以及为降低死亡率和流行病发病率所实施的公共健康项目。20世纪30年代,各学科的社会科学家开始致力于用严格的研究方法评估社会项目,由此,使得系统的评估活动变得越来越频繁(Freeman, 1977)。譬如,列文(Lewin)开创性地对"行为研究"的研究、李普特(Lippitt)和怀特(White)对民主和集权领导的研究,都是影响广泛的评估研究;类似的例子还有,涉及劳动生产率的著名的西部电器实验发现了所谓的"霍桑效应"(参见 Bernstein and Freeman, 1975,进一步的讨论参见 Bulmer, 1982; Cronbach et al., 1980;不同的历史视角看法参见 Madaus and Stufflebeam, 1989)。

从这些开创性研究开始,应用社会研究得到了加速发展,特别是第二次世界大战中,应用社会研究的贡献尤其明显。斯托福(Stouffer)及其同事和美国军队一起,获得了用于监测士兵士气的方法、评估人格的策略以及宣传技术。这些研究发现被美国战争情报机构用来监测军人的士气(Stouffer et al., 1949)。还有很多规模较小的研究,如评估价格控制和媒体攻势在改变美国人饮食习惯方面的功效。在英国和全世界其他地方的社会科学领域,也有类似的努力。

评估研究的繁荣期

第二次世界大战以后,为了满足城市发展和房产、技术及文化教育、职业培训、预防疾病等方面的需要,出现了无数由联邦政府或私人资助的发展项目。同一时期,还有大量由联邦政府机构或者私人基金支持的国际项目致力于家庭计划、健康和营养以及农村发展。由于花费巨大,项目各方理所当然要知道"结果如何"。

20世纪50年代末期,项目评估研究变得很流行,社会科学家们忙于从事防止过失、精神心理治疗和精神药物治疗、公共住宅、教育活动、社区组织创建以及无数其他类型项目的评估。这样的评估研究不仅发生在美国、欧洲和其他工业化国家,也发生在不太发达的国家。渐渐地,亚洲的家庭计划、拉丁美洲的营养与健康、非洲的农业与社区发展等类型的项目,都成为了评估研究的重要内容(Freeman, Rossi, and Wright, 1980; Levine et al., 1981)。同时,社会研究方法的发展和相关知识的普及,包括抽样调查、高级统计方法,以及不断增长的基金资助和行政管理知识,使得大规模、多层面的评估研究成为可能。

在20世纪60年代,评估研究论著的数量急剧增长。比如,海伊斯(Hayes, 1959)阐述了评估研究在欠发达国家的发展,萨奇曼(Suchman, 1967)对评估研究方法本身进行了回顾,坎贝尔(Campbell, 1969)举例说明了社会实验方法。在美国,引起人们对评估研究极大兴趣的,是由约翰逊(Lyndon Johnson)总统在联邦范围内发起的有关贫困的争论。到20世纪60年代后期,用华尔街的话说,评估研究已经变成了一个成长中的产业。

在20世纪70年代早期，评估研究已经成为了社会科学界的一个重要学术领域。各类书籍纷纷出笼，包括第一本教材(Weiss, 1972)、对评估研究方法论的批评(Bernstein & Freeman, 1975)以及涉及评估研究组织与机构局限的讨论(Reicken & Boruch, 1974)。1976年，《评估评论》(*Evaluation Review*)创刊，并逐步成为评估工作者广泛阅读的刊物。后来，又相继出现了不少期刊，到今天为止，评估方面的期刊已达十多种。在这一时期，许多学术和实践会议的热门话题就是评估研究。也是在这一时期，业界还成立了专门的评估研究者协会(专栏1—B列出了评估研究的主要期刊)。到1980年，科隆巴赫及其同事已经在说，"评估研究已经成为了美国社会科学中最有活力的前沿阵地"(Cronbach et al., 1980: 12-13)。

评估研究的发展也导致了评估研究本质的变迁。在早期，致力于评估研究的主要是社会研究者；但是到后来，评估研究的使用者(消费者)对这个领域也产生了重要的影响。现在，评估研究的持续发展则主要缘于决策者、项目计划者和行政管理人员的推动，他们利用评估研究的成果，并认为评估研究的成果值得信任。当然，评估研究的发展还得益于公众和项目对象的支持。尽管评估研究不能制造新闻，但是，评估研究的发现却总会牵动项目主办方、决策者、有见地的公民以及利益受到项目直接或间接牵连的人，影响范围十分广泛。

专栏1—B　评估研究的主要期刊和专业组织

主要致力于项目和政策评估的期刊

- Evaluation Review: A Journal of Applied Social Research (Sage Publications)
- American Journal of Evaluation (JAI Press)(formerly Evaluation Practice, before 1998)
- New Directions for Evaluation (Jossey-Bass)
- Evaluation: The International Journal of Theory, Research, and Practice (Sage Publications)
- Evaluation and Program Planning (Pergamon)
- Journal of Policy Analysis and Management (John Wiley)
- Canadian Journal of Program Evaluation (University of Calgary Press)
- Evaluation Journal of Australasia (Australasian Evaluation Society)
- Evaluation and Health Professions (Sage Publications)
- Educational Evaluation and Policy Analysis (American Educational Research Association)
- Assessment and Evaluation in Higher Education (Carfax Publishing Ltd.)

项目和政策评估的专业组织

- American Evaluation Association (Web page: http://www.eval.org/)
- Association for Public Policy Analysis and Management
- American Educational Research Association(Evaluation Division)
  (Web page: http://aera.net)
- Canadian Evaluation Association
  (Web page: http://www.unites.uqam.ca/ces/ces-sce.html)

- Australasian Evaluation Soceity (Web page:http:// www.parklane.com.au/aes/)
- European Evalution Society (Web page: http://www.europeanevaluation.org)
- UK Evaluation Society (Web page:http://www.evaluation.org.uk)
- German Evaluation Society (Web page: http://www.fal.de/tissen/geproval.htm)
- Italian Evaluatioan Society (Web page:http:// www.valutazione.it/)

消费者观点的介入,使得评估研究不再只限于让应用社会科学研究者对社会项目进行研究。同时,也使评估变成了政治和管理等复杂的活动。评估的目的就是,使政策决策、资源利用、项目设计、项目实施和延续变得更加有利于人的发展。在这样的意义上,评估研究应该是社会政策和公共行政运动的有机组成部分。

社会政策和公共行政运动

社会项目和评估活动出现的一个重要背景是晚近一段时期以来社会和环境状况以及国民生活素质等方面的责任向政府的转移。正如布伦纳(Bremner,1956)所描述的那样,第一次世界大战以前,除了战争老兵以外,公共服务基本上是个人和志愿组织的义务和责任。地方慈善组织善事活动的对象主要是穷人、残疾人和有麻烦的家庭。在我们头脑中,从事慈善活动的志愿者往往是这样一种形象——用篮子装上食品和衣物,施舍给那些不幸的人们。和民间团体、地方性慈善医院、县和州政府收容所、地方性公立学校、州师范学院以及养老院一起,志愿者们筑起了我们的公共服务"系统"。的确,20 世纪 30 年代以前,政府规模相对较小,尤其是联邦政府。譬如,20 世纪 20 年代,如果一年要花几十亿美元为老年人或穷人购买医疗设备和药品,就会使政府官员感到荒唐。联邦财政预算中,用于公共教育的支出也少得可怜——现在联邦政府用于公共教育的几个月支出就远远多于 20 世纪前十年的总和。

相应地,社会和经济信息的需求也很小。即使到了 20 世纪 30 年代末期,联邦支出中,每年用于社会科学研究和统计研究的份额,也只有 4 000 ~ 5 000 万美元,而今天的份额已经是那时的很多倍。同时,那时的公共服务和政府运作规则也与今天的不相同。重要政府官员的选择往往不考虑客观的竞争性标准;事实上,那时候用来评定个人能力的客观方法几乎没有。那时,专业的公共服务仅仅相当于现有总量的很小一部分,大多数的公共服务工作都不需要专门的技能;同时,服务人员也没有受过什么正规的训练。

所有这些,在 20 世纪 30 年代以来都得到了改变。随着大萧条的到来,公共服务获得了飞速发展,当然,政府部门也是如此。一部分原因是,公共项目的快速发展,形成了应用科学管理理念和技术的强大压力;特别是科学管理在工业领域已经有了很好的应用,给政府项目和活动的压力自然就很大。首先采用科学管理理念的是国防部,接下来就在政府其他组织和部门扩散开来,包括社会服务机构。类似于计划、预算、质量控制、责任以及后来更加复杂的如成本—收益分

析和系统控制模型等，逐步变成了人们日常工作的一部分。

政策研究和公共行政专家

在同一个时期，接受过社会科学训练的人，开始把自己所学到的知识运用到理解实践部门和机构的政治决策，并运用到组织和行政决策中。这个时候，在某种程度上，社会科学家对政府的兴趣完全是学术性的——他们希望知道政府是如何运作的。但是，在政府领导岗位上的人，摸索着如何管理大量的人力和资金，特别需要使工作有条不紊，并由此细化政策、行政、项目和计划的责任。他们逐步认识到，经济学、政治科学、心理学和社会学中的概念、技术和原理，也是有用的。随之，对公共部门的研究，便基本上成为了应用研究，这样的研究现在被称为“政策科学”和“政策分析”。

随着联邦政府机构变得日益复杂化和技术化，政府雇用的智囊已经远远不能满足政府项目的需要，或者是因为这些智囊人员与政府的关系（如同党、亲戚、朋友）而不能让他们管理项目。由此，需要大量的、接受过专业训练、具有相关技能或丰富经验，且具有竞争力的中层管理人员和许多高层主管（参见专栏 1—C）。作为对这种需求的回应，很多大学的管理、公共卫生和社会工作研究生院，通过一些培养项目向政府部门输送大量的经过训练的人才。更多的专门的职业技术学院不断建立或者扩展，开始为满足对主管类和技术类人才的需求而开设和扩展专门的研究生课程——公共行政。

今天，对评估重要性的认识，已经成为政府官员和高级管理人员的共识。此外，许多联邦政府机构都已有自己的评估部门，联邦政府在各州的派出机构也如此。此外，联邦、州和地方都把评估工作用合同的方式交给大学研究人员、研究机构、咨询机构，这已经成为了惯例。总之，尽管评估研究仍然有其学术的一面，包括培训、方法、理论以及社会项目的特点和效果；但是，作为一个领域，评估研究已经远远超出了学院式社会科学的围墙。在政策制定、项目管理、为客户辩护等方面，评估已经变成了一种日常工作实践。因此，不仅评估研究的发展史涉及社会政策和公共行政运动，而且评估研究的实践也主要发生在政策分析和公共行政等政治和组织领域。

专栏 1—C　政策分析的兴起

政府面临的政策问题数量日益增加、复杂性和变数日益增长、社会重要性日渐突出，使得政府对具有这类知识的官员和职员的需求日增。面对核安全、少女怀孕、城市衰落、日益增加的医疗成本、年轻黑人失业率的上升、家庭暴力的增长、有毒废弃物的处理等问题，到底应该怎么办？20 年前，许多这样的问题都不曾表现为公共性议题，而现在却变成了紧迫的社会难题，此外每年还有许多新问题出现。对任何当选和被任命的官员及其职员而言，如此复杂的和具有争议性的问题，已经超出了他们所能够驾驭的范围。但是，又不能把这些问题抛在一边不管，政府高级官员期望能够对这些问题进行有效的回应和处理。

为了帮助政府官员思考和处理这些问题，就需要从政策分析、项目评估以及统计资料中，不断

地吸收新的知识、扩大视野。与过去不同的是,现在的官员(无论是当选的还是被任命的),在决策和处理问题时,常常会引用已有的研究成果、官方数据和专家观点。一般的政府职员数量已经有了很大增长,所承担的责任也相应增加,他们之中接受过专业训练和熟悉分析技术、懂得搜集、甄别信息的人的比例也提高了。大量的研究、分析和资料搜集工作正由他们来完成。

由于政策的影响力直接涉及政府系统,所以,政府官员如果要影响政策(使政策能够更好地实施),就必须使其要有说服力。由于政策问题的特征不断改变,所以,获得说服力就变得更加困难。资历、亲和力、聪明伶俐等特征的影响力日弱,知识、思考和处理问题的见解以及具有充分信息源等特征的影响力日强。渐渐地,从总统以至下层官员,在政策辩论中,如果不能提出正确的数据或提出的数据被对方的专家驳倒,就会失去影响力。的确,在某个问题上,详细透彻的指令常常是强制性的。而获得合法性,则常常要求主管本人就是自己工作领域的专家。公正性则要求详细地证明,行政决策不是随意的和反复无常的。预算官员需要的是对项目正面的评估,公众需要的责任承担。这样,从政治体系面对社会问题的动态过程可以看出,其在公共事务方面提高了实际处理能力和管理能力。

资料来源:Laurence E. Lynn, Jr., *Designing Public Policy* (Santa Monica, CA: Scott, Foresman, 1980).

从美好社会运动到当前的评估事业

在肯尼迪和约翰逊当政的20世纪60年代,打着"反贫困战争—美好社会运动"的旗子,大量社会项目获得实施,评估活动迅速成长。"反贫困战争和美好社会运动"的主要目的就是提供大量资源,解决失业、犯罪、城市衰落、医疗供应短缺、精神疾病治疗困难等问题(参见专栏1—D)。这些项目的上马,往往比较仓促、设计不足、实施不当、管理低效。大量联邦政府项目的低效和低投资回报率,使得政府不得不考虑对社会项目的长远效果进行重新评价。

专栏1—D　20世纪60年代政策分析和评估研究的成长

1965年,"政策分析和评估研究"变成了研究领域的一个独立分支,这是一个重要的时点。在联邦政府方面,有两项发展非常重要,"反贫困战争—美好社会运动"和"项目规划预算系统"实施法令。这两项发展不仅为学者的活动提供了正当性支持,也提供了合法性和经费支持,使学者愿意把他们的研究兴趣转向考察公共资源的效率、对个体行为的影响、项目设计的有效性以及缓和贫富差距、种族和少数种族矛盾以及南北矛盾的效果。

1965年发起的"反贫困战争—美好社会运动",代表了一种规模空前的社会干预活动。所有受到影响的人都想知道:如果他们投入这一工作,谁会受到影响,以及如何受到影响。而有能力回答这些问题的人,不仅需要经费支持,而且要有受众,这就是社会科学的学者们。也是在这一年,政府部门广泛地采用了正规的评估与分析方法;而在过去,只有在麦克纳马拉(McNamara)时期国防部的"项目规划预算系统"中才用过。一项总统的实施法令,使得那些希望把自己这类才能贡献出来的人不仅有了职位,而且有了经费支持,以便使自己能够回答诸如效率、绩效和平等之类的问题。

资料来源:Robert H. Haveman, "Policy Analysis and Evaluation Research After Twenty Years," *Policy Studies Journal*, 1987, 16:191-218.

到20世纪70年代，在一定程度上，因为许多政府项目的低效，而且项目费用不断增长，使得社会项目面临着严重的财政问题，社会上对继续扩大政府项目的抵制力量日益增长(Freeman & Solomon, 1979)。由此，也带来了评估领域的变化。评估的关注点，更多地转向了社会项目的成本而不是收益、财政能力和有效管理。在这个过程中，许多财政和政治保守主义者，特别是对社会科学持怀疑态度的人，加入到鼓吹迫切需要项目评估提供信息支持的社会项目的阵营。

到20世纪80年代，从里根政府开始，一直持续到现在，为控制通货膨胀和降低联邦赤字，政府大量缩减了联邦政府的国内经费。其中，最大的项目缩减都集中在社会项目方面。许多州和城市也采用了同样的策略。的确，许多州和城市对其经济不景气的反应更加激烈。部分原因是由于一些社会成员对收入和财产税政策感到沮丧，进而采取不信任、敌意的政治行动，从而也导致了经费的缩减。但是，正如我们已经指出的，这样的状况同时也是过去几十年中一些政府官员、计划人员、政客等随意设置社会项目、项目实施不力、项目效果低下所造成的。

很明显，社会项目(包括评估事业)强烈地受到这种时代变迁的影响。近几十年来的政治视角，不仅在美国，也在西方国家，使得人们更多地关注社会项目的成本—收益的平衡。在学术界前沿，对“美好社会运动”项目保守的和自由的批评者，对评估领域都产生了影响。尽管在某种意义上，这些批评更多地表现为意识形态，并缺乏足够的证据。但是，这些人却利用评估的结果，来谴责社会项目。正因为如此，评估研究被夹在了社会干预问题争论的中间而面临这样的格局，那就是要极力证明任何有创意的项目都可能产生良好的效果。

当我们进入21世纪时，财政保守主义、向州一级政府的责任转嫁以及对社会项目的怀疑主义主导着国家层面的政策制定。这些趋势和政治走向影响着评估的应用与发展。一方面，如果要了解这些改革的财政和社会影响，社会项目方面对这些重要修订和改革要求进行评估，事实上，有大量重要的国家评估项目都仍在实施之中(Rossi, 2001)。另一方面，随着项目被转移到了州政府层面，进行评估的责任落空了。尽管州政府或当地主导的项目评估数量和质量都有了稳定的上升，但是许多州政府仍没有能力或者不愿意提升这种能力来进行严肃而有效的项目评估。

无论政治发展趋势如何，当前评估环境的两个方面已经非常清楚：第一，对资源的紧缩会继续要求慎重地选择需要优先考虑的社会问题项目。第二，因为要求缩减成效不大的项目的压力仍然存在，对既有项目的慎重审查会继续下去。此外，对既有项目的不满和政治气候的变迁，都会迫使新的和修订的项目承诺更加高效和节省。所有这些因素，都会对评估研究产生重要影响。

项目评估的特征

当我们对评估研究的历史已经有了一些了解以后,就可以在当前背景下,给项目评估下一个比稍前更完整、更适用的定义了。作为一本教科书,我们必须从一个恰当的定义开始:

项目评估就是采用社会研究的程序,在一定的政治和组织环境下,系统地调查旨在改善社会环境和条件的社会干预项目的绩效。如果对这个定义中的主要部分进行系统地解释,就会使我们对本书讨论的项目评估有一个基本了解,使我们相信项目评估的主要论题是整合于评估实践过程之中的。

社会研究方法的应用

一方面,评估是指对整个被评估事物绩效的描述;另一方面,还涉及用来说明绩效的标准和测量(参见专栏 1—E)。也就是说,项目评估者的中心任务就是,建构与项目绩效有关的、可以与一定的标准进行比较的、有效的描述。如果不能在一定效度层面上表述项目的绩效,就会扭曲项目实现的内容、否认项目成功的希望,或者过分看重项目的缺陷。而且,这样的描述必须是细节性的和准确的。对项目绩效进行不适当地褒奖或者不真实的描述,也会使人们对项目绩效是否符合某些实际标准产生怀疑。

在过去的时间里,在对社会现象进行有效描述方面,**社会研究方法**以及与之相关的方法质量控制标准已经得到了很好的发展。特别是,当代社会科学中的系统观察、测量、抽样、研究设计和数据分析技术,已经发展到了相当高的水平,已经能够对社会行为的特征进行有效的、可靠的、准确的表述。因为社会项目是有组织的社会行为,所以不证自明的是,从有效和可靠的角度来看,社会研究方法为描述社会项目的绩效提供了最好的方式。

不管我们所研究的社会干预属于何种类型,评估研究者将主要使用社会研究方法搜集项目绩效的资料,并对资料进行分析和解释。对社会研究规范的遵守,是评估研究的核心观点,也是本书副标题"系统方法"的含义。但这并不是说,评估研究一定要遵循社会研究的某种模式或者某几种模式的结合形式,无论这些形式是定量的或定性的、实地的或民族志的、"实证主义的"或"自然主义的"。同样,评估研究对社会科学方法的遵从,也不意味着现有的方法已经达到了极限,再也没有改进的余地。在获得可信和有效的项目资料方面,评估研究者必须不断创新和改良。实际上,评估研究者已经并将继续在应用社会科学方法发展方面贡献自己的力量。

专栏1—1 评估研究的两个重要方面

评估就是确定事物优点和价值的过程，对事物的评估就是这个过程下的结果……所以评估并不只是简单地搜集和整理与决策有明确关系的资料，尽管现在仍然有评估专家这样定义评估……在任何情况下，对决策而言，搜集和分析资料（通常也是很困难的）都是必须的，但这只是评估工作两个重要部分之一，如果缺失了另一部分，即对资料的综合，也就称不上什么评估了。《对象报告》并不只是检验产品、报告检验的分数，而是要①依据功效或成本—收益，并②计算比例和排名。要获得这样的结论，在大多数情况下，需要投入除了数据以外的东西。获得结论的第二项要求就是，确定功绩或净收益、可评估的标准……更简单地说，评估有两个重要方面：一方面是搜集资料，另一方面是整理和鉴别资料的相对价值和标准。

资料来源：Michael Scriven, *Evaluation Thesaurus*, 4th ed. (Newbury Park, CA: Sage, 1991), pp. 1, 4-5.

最后，上述观点并不意味着方法意义上的质量是评估的最重要方面，也不意味着在评估中只有采用最高的技术标准而不实行妥协才是好的。正如维思（Weiss, 1972）曾经观察到的，社会项目天生就具有不适合做研究的环境。项目环境的特质（特别是要求评估者关注的议题），常常与教科书中的标准有一定的偏差。具体项目的操作环境、评估者必须面对和回答的具体问题，经常驱使评估者与环境相妥协，并对教科书中的方法和标准进行调整。因此，对评估者而言，最大的挑战就是，将研究方法和程序与所需要回答的问题及项目环境进行具体匹配。也就是，如何采用学术的程序来组织评估的问题和程序，并让评估研究得以实施。无论用什么方法，都要采用最合适的标准，使研究方法对于问题和环境具体可行。

社会项目的绩效

我们把社会项目定义为“做好事”，这是其得以延续的最重要原因。也就是说，通过解决社会问题或者改善社会状况和生活水平。投资方因为社会项目对于社会改善的贡献而坚持对项目的资助。当然，对任何值得评估的项目都要进行评估。也就是说，必须要对项目的一个或几个方面进行评估。更具体地讲，项目评估涉及下述五个项目维度的一个或几个方面：①对项目的需求，②项目的设计，③项目的实施，④项目的影响或产出，⑤项目的效率。在后续章节中，我们将展示评估者如何开展以上几个维度上的评估。

评估计划针对的是一系列的问题，这些问题一般由**项目主办方**（Evaluation sponsor）提出；而另一些**项目方**（Stakeholder），不管是个人，还是群体、组织，对于项目问题的确定和项目功能的良好发挥都有重要的影响。当把这些问题具体表达出来的时候就构成了评估服务的细节性协议。而且，评估者必须协调项目主办方和其他项目方，提炼和深化问题。因为，尽管这些人很可能知道自己的利益需求和项目目的，但是他们并不必然用评估者熟悉的标准和概念来表述他们所关心的问题。

在特定的政治和组织背景下进行评估

项目评估并不是快刀斩乱麻式的活动,也不像搭积木或者用文字处理程序检查拼写错误。相反,评估是一种实践。在设计评估方案时,必须根据特定的项目环境进行调试,在评估实施中还要不断地修订和修改。某项评估的特定范围和形式,主要取决于评估的目的和评估结果的受众、被评估项目的性质以及评估实施的政治和行政环境。

譬如,初始问题也许仅停留在项目口号层面,如果要评估,就要把空泛的、口号性的问题转化为具体的问题。偶尔,评估问题也会是预设性的(譬如项目绩效),并不直接来自对相关议题的深思熟虑。在这种情况下,评估者必须细致地厘清,这些含糊不清的问题对评估各方到底意味着什么以及他们为什么关心这些问题。

与评估问题确定同样重要的是,要弄清楚为什么会问这些问题,回答这些问题又会有什么意义。项目评估必须提供所关心主题的信息并以及时而有意义的方式向决策者不断反馈信息,同时,以一种适用于这些目的的方式与决策者等主体交流,例如在设计上,一个旨在给项目经理反馈信息并通过提高服务质量以改善项目的评估和一个旨在反馈信息给项目资助者以便使其决定是否继续资助该项目的评估,会有很大的差别。不过,在任何情况下,评估的规划和实施都要慎重考虑其中的政治问题(参见专栏 1—F)。

以上观点假定除非受众愿意接收评估结果、大众至少会潜在地使用这些评估成果,否则评估活动无法开展。不幸的是,项目主办方有时候并不是出于实际应用的意图而开展评估研究。譬如,一项评估会被项目资助方操纵,以至于项目评估屈服于资助方的需求。有责任心的评估者必须努力避免陷入"仪式性评估"的境地。因此,计划和开展一项评估的第一步,是全面透彻地询问项目主办方的动机、评估的主要意图以及如何使用评估结果。

当然,作为一项实践活动,评估还必须考虑项目的组织层面。在进行评估设计的时候,必须考虑到,行政上的合作与支持,项目文件和资料的可及性,项目服务的特征、项目与客户之间关联的性质、频率、时间间隔和范围。的确,评估活动一旦展开,进行调整和修订也是非常常见的事情。由于不可预见的困难和政治障碍,在必需资料的类别、数量或质量方面会出现调整,甚至妥协,这都是常有的事情。此外,如果项目方的利益构成和操作方式出现重要改变,那么就要对评估的基本问题进行调整。

通过知会社会行动来改善社会环境

正如已经强调的那样,项目评估的角色就是向人们提供答案。它所要回答的问题就是项目怎样做才是有价值的,项目实际上又如何被实施。对评估研究而言,这是最基本的,即评估研究的基本目标就是知会社会行动。因此,对评估研究特别关注的,就是那些根据评估结果来决策和采取行动的各类对象。评估

的发现也要有助于决策者的决策，譬如对项目的某个方面进行调整，或者启动新的项目，或全部继续已有的项目。这些人要考虑政治的、实践的和资源的方方面面，或者获得对项目对象的基本印象。也许，他们能够直接判断项目是否能达到目标。或者，他们能对项目的形式与构成以及与此有关的争论产生间接影响。

专栏1—1 评估中的政治诉求

评估是一定政治背景下的理性活动。政治考虑因素的介入主要通过三种方式，评估者如果认识不到这些，就会深受困扰：

第一，要求进行评估的项目和政策，都是政治决策的产物。这些项目和政策的设计、界定、争论、实施与资助，都在政治过程中完成，并受到来自各方的压力（支持的、反对的），这就使得政治问题非常重要。

第二，由于评估是为了满足决策的要求，所以评估报告要进入决策领域。在政治过程中，项目产出的评估性证据要有足够的竞争力。

第三，也是最没有引起重视的就是，评估本身就具有政治意涵。就评估本身的特点而言，它使得针对某些项目的问题变成了绝对的政治陈述，并使其他项目无法对其提出挑战，也使得项目目标和战略的合法性、渐进改革的战略用途，甚至社会科学家在政策和项目形成中的角色，都变成了政治性问题。

知晓这样的政治局限和对抗性力量，并不构成放弃评估研究的理由；相反，是进行有用评估研究的前提。评估者只有洞悉系统中其他社会行动者的动机和利益、评估者的角色、完成评估的机会和障碍以及应用评估结果的限制和可能途径（特别是在政治上敏感的评估），才能使得评估变成创造性的、战略性的和有用的活动。

资料来源：Carol H. Weiss, "Where Politics and Evaluation Research Meet," *Evaluation Practice*, 1993, 14(1): 94, where the original 1973 version was reprinted as one of the classics in the evaluation field.

项目，就和人一样，具有自己独有的特质，当然也有和其他项目共有的、具有类群意义的特质。就具体项目而言，譬如在某中学实施的预防药物滥用项目，评估不仅会告诉我们在特定高中该项目的实施状况，而且会告诉我们同类项目的状况。卷入社会干预的利益集团，主要是针对某些类型的项目采取行动，而不是针对个别的项目。联邦立法机构也许考虑为教育项目追加投资，州政府机构也许考虑预防青少年犯罪的项目，慈善基金会也许鼓励让护理人员造访单身母亲的项目。对于这些项目，对于这些类型的决策和社会行动，评估的发现是至关重要的。

的确，评估研究的一种重要形式就是示范性项目，即通过细致的设计、实施社会干预项目来检验创新性项目的价值。在这种情况下，因为对项目的概念性评估会比任何其他类型的评估更加直接地影响到政策制定和项目发展，所以，评估研究的发现就特别重要。另一个和评估相关的重要活动就是对各种评估发现进行综合、进而知会政策制定和项目规划。

因此，评估可以通过为计划和政策制定提供信息，指明针对社区问题所采取

的应对方法是否值得,展示专业实践中特定原则的有用性,来知会社会行动评估研究甚至还会通过检验某种广泛干预形式的社会科学假设,来帮助我们理解如何达成有计划的社会变迁。一般而言,评估研究是有用的,而且被广泛地应用着,并因此直接和间接地为实践知识的积累做出贡献。

实践中的评估研究

我们已经解释了影响评估研究的一般性考虑、目标和策略。在实践中,这些概念的应用,还要涉及各种力量的平衡。在这些关系中,极为重要的一面就是其中固有的冲突,譬如系统化的要求和根据评估需要进行资料搜集时的冲突、提供某些服务的组织强制与日常工作维护的冲突。所以,在评估的准备阶段,就要协调好项目的人事和相关安排,这样,在资料搜集阶段,就能够获得各方的支持。譬如,资料搜集就会涉及获取项目文件、客户及工作人员支持;以及处理不好这些问题就会打乱项目正常进程,分散甚至危及项目的基本责任履行。

因此,每个评估规划必须尽量做到这一点:既不对评估研究的环境过分乐观,也不对评估活动引起的干扰性影响过分保守,也就是要进行折中。在这里我们慎重地使用"折中"这一概念,因为处理评估研究与项目运作之间紧张关系的最好方法,就是评估者与被评估者共同制订评估规划。在实施研究之前,如果能够详细厘清评估的需求和目标,那么,被评估者(不仅仅是管理者)就有机会对评估本身做出反应,并对此做出贡献(譬如资料采集),也能够得到一个比较适用的工作方案,并能够在实际评估活动中得到被评估者的支持,缓解评估者与被评估者之间的紧张关系。

除了评估活动与项目活动之间的冲突以外,还有另外的与评估活动有关的冲突。这里,我们再介绍一些评估者必须面对的困境:既有评估方案与社会项目多变性之间的不兼容;按科学研究要求的评估与实际操作之间的差异;因评估对象的多样性所造成的方法之间的不兼容。

评估与社会项目的多变性

项目评估最具挑战性的一面,就是被评估社会项目决策的不断变化,特别是资源、优先秩序、项目各方对项目的相对影响是动态的。而与这些变化伴随的,常常是我们前面讨论过的政治背景和社会形势的变动。譬如,1996 年的福利制度改革就直接触及对贫困家庭的支持。对项目的重新配置,就要求直接涉及与过去不同的对贫困家庭支持的评估,由此就会涉及新的、对项目产出的评估和不同的项目内容评估。

在这种情况下,涉及项目实施机构的优先秩序和责任划分,也会发生明显的变化。譬如,某学校虽然在强制实施校车制度方面得到了帮助,但是,在主要是黑人的学校中增加白人学生的入学率却可能损失利益。或者随社会干预出现

的、不可预见的问题,会要求修改项目,并因此影响评估计划。举例来说,一个旨在通过提供医疗保障来减少低收入家庭中学生失学率的项目,会因为大多数项目对象拒绝接受项目服务而受到阻碍。

具有讽刺意味的是,评估研究的初步发现也会促进项目的变化,并因此使接下来的评估活动陷入困境。譬如戒酒项目的影响评估包括未来6个月和12个月的观察。当6个月的观察显示出较高的醉酒率的时候,就要修改项目方案了。

无论如何,评估研究者必须事先考虑到这样的变化,并尽可能地做好准备。也许,更为重要的是,要使评估活动与项目环境和评估时点规划一致。在项目受到决策各方的影响而进行重要修订的前提下,想要制订完美的影响评估方案是不切实际的。同样重要的是,在评估过程中,评估研究者针对评估活动所展现的弹性。如果评估规划已经明显地与环境不适,就要了解项目的动态性,评估研究者也就必须为在评估过程中的修订做好准备。这种情况通常是因为缺乏与评估活动有关的资源、时间紧迫、评估研究者与被评估者的关系不明确而变得紧张,所以,这样的困难往往不容易解决。尽管社会项目不是实验室,但是,评估研究者必须为自己能力以外的力量和事务变化做好准备。

在评估研究中,实验室与社会项目作为研究场所的反差,使得评估研究者不得不面临另一个问题,即科学的与实用的取向之间的内在张力。

科学的与实用的评估态度

也许,在评估领域最有影响的是坎贝尔(Compbell, 1969)的文章。这篇文章聚集了坎贝尔几十年的心血,以表达这样一个观点:政策与项目决策应该来自于不断的、旨在改善社会条件的社会实验。他不仅坚持这样的立场,而且致力于发展这样的社会研究技术,以使真正的"实验社会"成为可能。因此,坎贝尔也试图发展实验模型,并将其应用于社会心理研究和评估研究。尽管在后来的写作中他调整了自己的立场,但是,我们还是应该把他纳入具有科学研究范式的评估研究者行列(参见专栏1—G)。

专栏1—G 作为实验的改革

美国和其他现代国家应该采用实验取向来进行社会改革,即尝试用新的社会项目来解决社会问题,根据多种标准来判断各种项目的绩效,并了解到这些项目是否有效,进而决定保留、模仿、修订或者放弃。

资料来源:Donald Campbell,"Reforms as Experiments," *American Psychologist*, April 1969, 24: 409.

在评估研究领域,坎贝尔的立场受到了另一位大师科隆巴赫(Cronbach)的挑战。尽管他了解评估研究要采用科学的调查并在评估中用到科学逻辑,但是,科隆巴赫认为,评估的目的已经把评估研究与严格的科学研究进行了区分(Cronbach, 1982)。在他看来,评估比科学具有更多的艺术成分,每一项评估都要进行调试,以适合项目决策者和相关方的需要。因此,尽管科学研究要求满足

研究标准,但是评估研究却要在既有的政治环境、项目局限和可用资源下,最大限度地为决策者提供有用信息(参见专栏1—H)。

人们也许倾向于同意这样两种看法,评估应该满足高质量的科学研究标准,同时也应该为决策者提供决策所需的充分信息。当然,问题是在实践中,这两种目标常常相互不兼容。特别是,高标准的社会科学研究需要的资源常常超出一般项目评估的范围。这些资源包括时间(因为高质量的社会科学研究不可能是急就章,而项目决策常常需要在较短的时间内做出)、经费、努力的程度以及设备。此外,在科学框架内的研究必须具有研究结构,而评估研究中,不一定有这样的结构。譬如,用于科学研究的变量需要具体地定义和测量,而这样的做法会使得决策者认为,把复杂的和动态的项目变得支离破碎从而难以捉摸。同样,在调查项目产出时(项目是否是可观察到的变化的原因),要满足科研中的因果关系,就必须要有实验控制(即对照),而这却不属于项目服务的范畴,有时还在一定程度上干扰、限制实际的项目操作。

专栏1—H 作为教员的评估者

在某种程度上,对社会项目的评估研究就是判断项目的政治意涵。所以,对项目的评估,就在于判断项目为大众利益做出的贡献以及服务的质量。评估的价值就在于,让人们对某些行动和结果了解得更清楚。所以,评估本身不仅仅只是提供一堆数据,而是要及时(不是最后)地沟通,为关注项目的人们提供信息,并使获得信息的人把信息纳入自己的思考。更广义地说,评估一定要知会和改善社会系统的运作。

资料来源:Lee J. Cronbach and Associates, *Toward Reform of Program Evaluation* (San Francisco: Jossey-Bass, 1980), pp. 65-66.

另一方面,人们在评估研究中又不可能不顾及科研标准。确切地说,科研方法所代表的是使评估结果有效和可信的取向。即使有时候达不到这样的要求,这样的努力也为决策过程做出了重要的贡献,尤其是在避免个人利益、武断、意识形态偏见、私下交易等方面做出了巨大的贡献。不过,这样说有一个前提,那就是评估研究为决策者提供了有意义的东西,否则这样的研究就没有信度和效度;只是,这里的信度和效度与科研的信度和效度无关。

因此,在实际工作中,评估者始终要在强调工作程序,确保在使结果具有信度和效度与促使评估发现及时获取并对客户有意义两个方面达成平衡。而平衡点则完全取决于评估的目的、项目的性质以及政治和决策的背景。尽管评估研究中的具体限制(项目条件和可用资源)不可能使评估研究变成设计良好的科学研究,但在许多情况下,评估应该能足够好地回答相关的政策和项目问题。

如果在各类评估结果的用户识别以及潜在用户的优先秩序等问题上含混不清,那么,就会为评估规划增加更多的困难。一项评估通常有多种类型的潜在受众,其中有的与被评估项目的某些部分有直接利益关系,也有的只是对项目所代表的社会干预类型感兴趣,还有的也许处在两者之间。有时候,事先就已经把评

估的目标和优先用户定义得很清楚，这会使得评估研究者在平衡科学与实用取向时少一些困难。但是，许多项目评估的状况并不是这样明确。评估也许只是作为日常项目经费和合同管理的一部分，并因此假设评估活动会为管理者、资助者以及其他相关方面提供信息。或者评估活动是一种合作：服务机构需要信息以改进管理；研究者需要广泛地了解特定项目所展现的社会干预。的确，评估活动的目的往往是多样性的、分散的，而不仅仅是把各种目的紧密地结合在一起。在这种情况下，项目决策机构对评估的运用和科学研究对评估的要求之间，很少有可能达成一致。

一些评估理论家认为，**评估的利用**（Utilization of evaluation）是最重要的考虑，并致力于与用户密切合作，进而设计符合需要的评估（Patton，1997）。相对应，应用研究期刊的一些作者则希望对各种干预的绩效进行整合；他们不断地对方法上有缺陷的评估研究进行批评，并倡导高标准。有些评论者希望评估研究能够两全其美，既要为项目各方所用，又能够积累社会干预方面的知识（Lipsey，1997）。我们的观点，特别是这本书的观点是，所有这些都有可能，但是在具体的评估中，可能性的大小却不一样。这就提出了另外一个问题，那就是评估者自己一定要进行判断，并为具体的评估目标和每种具体的环境进行规划调试。

评估观点和取向的多样性

正如前面的讨论已经说明的那样，评估研究领域在观点和方法方面并不统一。相反，评估始终是一个多样化的领域。坎贝尔和科隆巴赫代表了历史上两种基本不同的取向，但也仅仅只是这种差异的一例而已。评估研究者来自不同的学术领域、具有不同的方法论背景，这样，多样化的学术交叉为评估研究带来了丰富的视角。观点不同的另一个来源就是评估者的动机和评估者工作环境的差异。临时从事评估工作且只是进行地方性小型评估的评估者和长期从事评估工作的专职人员，对评估活动的观点会有相当大的差异。

随着评估领域的成熟和制度化，人们凭其在评估中的兴趣和方法也被细化为不同的阵营。特别是，人们有越来越多的兴趣从不同视角集合相似的观点出发来发展所谓的“评估理论”（Shadish，Cook，& Leviton，1991）。致力于发展评估理论的运动使评估者在自己的工作中把理论应用于评估，进而成为决策的基础（参见专栏1—1）。

专栏1—1 理想的评估理论

理想（从来没有实现的）的评估理论应该能够描述和评判为什么某种评估实践能够在评估者面对的环境条件下获得那样特定的结果。它应该①能够厘清评估的活动、过程和目标；②厘清评估活动、过程和目标之间的关系；③经验性地检验本研究或其他研究之间相冲突的命题。

资料来源：William R. Shadish，Thomas D. Cook，and Laura C. Leviton，*Foundations of Program Evaluation：Theories of Practice*（Newbury Park，CA：Sage，1991），pp. 30-31.

所以,现在我们必须懂得,在评估中既有科学的部分,也有艺术的部分,也许评估应该如此,并将会继续如此。无论如何,评估者的任务就是创造性地把各种不同的关注和目标集合到一起,让不同的观察者能够从中找到不同的信息。当然,我们知道通过文字的方式来教授艺术的东西是很困难的。教授评估有点类似培训医生。任何有知识的人都可以通过训练而懂得实验室实验的结果,但是,良好的医生则只能通过临床、经历以及对每个案例特性的体验来塑造。从这种意义上说,在书本中,我们只能学到一个训练有素的评估专家的基本知识。

谁能从事评估工作

系统评估的基础是社会科学的研究技术。因此,评估专家必须接受社会科学的基本训练。但是,我们应该马上指出的是,在从事评估的队伍中,人们的专业背景具有极大的异质性(参见专栏 1—J)。理想上,每个从事评估研究的人都应该具备社会科学研究的基本知识。但是,我们也要知道,熟悉和了解任务领域的问题(譬如犯罪、卫生、吸毒)、性质、范围、以前所运用过的干预的评估结果等,也是非常重要的。首先,评估者必须明白项目所要处理的问题和项目开展的社会背景;另一方面,评估者必须根据项目的实际情况和类似项目的已有知识,制订出一个合理的评估规划。

在最复杂的层次上,评估活动在技术上、概念上可以变得异常复杂,也可以变得非常耗费成本,甚至需要训练有素的社会科学家(至少了解社会理论、资料搜集方法和统计技术)有相当长时间的参与。这样复杂的评估,常常由职业的评估队伍来完成。另一种极端的情况就是,存在许多比较容易完成的评估工作,这些工作甚至可以由经验不足的评估人员来完成。

本书的目的就是为那些现在从事评估工作、具有职业兴趣或者对评估好奇,并希望知道评估是怎么回事的人介绍评估领域。当然,对本书的学习只是成为职业评估人员的开始,并不能代替实际评估经验本身。这本书的另一个目的就是为那些从事人力资源管理的人员提供知识,为他们提供有关评估的概念和方法,让他们能够充分地理解评估任务和活动能够在何种程度满足他们对项目和计划的评判,并有能力理解针对其项目和计划的评估结果。简而言之,我们希望提供一本教科书,帮助那些从事评估的、对评估有责任的、经常接触评估的和评估研究的消费者理解评估。

专栏 1—J 美国评估协会会员的多样性(百分比)

按职业划分	2003
评估	39
研究	15
行政	10
教学	8
咨询	10
学生	6
其他	4
未知	9

按组织划分	2003
学院或大学	36
私营企业	18
非营利组织	17
联邦政府机构	7
州或地方政府机构	6
学校系统	3
其他	6
未知	8

按学科划分	1993
教育	22
心理	18
评估	14
统计方法	10
社会学	6
经济学和政治科学	6
组织发展	3
其他	21

资料来源:2003 年的资料是基于 3 429 个成员统计出来的,具体报告由美国评估协会(AEA,2003,4-19)的基斯特勒(Susan Kistler)提供;1993 年的资料源于《评估实践新闻》(*Evaluation Practice News*,1993 年 10 月),其数据是基于 1993 年 6 月美国评估协会的 2 045 个成员得出的。

小 结

- 项目评估就是运用社会科学方法,系统地调查社会干预项目的绩效。评估需要运用社会科学的概念和技术,以期对项目的改善有用,并以期通过知会社会行动来减少社会问题。
- 现代评估研究的发展起始于 20 世纪 30 年代,二战之后,作为新的研究方法被广泛地运用于迅速扩散的社会问题领域。社会政策和公共行政运动为评估的职业化做出了巨大的贡献,并使得评估研究的消费者复杂化。
- 项目评估的需求在我们这个时代并没有减少,甚至趋向于增加。的确,人们对稀缺资源的关注使得评估变得更为重要,甚至比社会干预的绩效更为重要。
- 项目评估需要对项目绩效或者所要评估的问题进行精确的描述,并依据相关的评判标准,对其进行评定。
- 一般情况下,评估工作会涉及项目的五个方面:①项目需求;②项目设计;③项目实施和服务发展;④项目的影响或结果;⑤项目的绩效。评估需要对项目的绩效或特征有准确的描述,并根据一定的标准对其进行评价。
- 在实践中,评估者会面临许多挑战。在评估期间,项目的环境和行动也许会改变;在评估设计中,还要平衡科学研究和应用研究的关系;而评估领域极其多样化的取向和视角,也不可能为我们提供所谓最好的评估流程。
- 大多数评估研究者都会在社会科学的一个或几个领域接受应用社会科学的训练。特别专门化、技术化或复杂的评估,需要职业评估人员的操作。但是,评估领域的基本知识不仅与从事评估的人有关,也与评估研究的消费者有关。

基本概念

评估主办方(Evaluation sponsor):要求进行评估并提供评估资源的个体、群体或组织。

项目评估(Program evaluation):运用社会研究程序,在一定政治和组织环境条件下,系统地调查致力于改善社会环境的社会干预项目的绩效。

社会项目;社会干预(Social program; Social intervention):有组织、有计划、持续不断地致力于解决社会问题或改善社会环境的努力。

社会研究方法(Social research methods):社会科学家基于系统观察和逻辑规则并依据观察所得进行推论的、研究社会行为的程序。

项目各方(Stakeholders):利益直接受到项目运作及绩效状况影响的个体、群体和组织(例如,对项目有权威性影响的个体、群体和组织,资助者和主办方,管理者和职员,客户或潜在的受益者)。

评估的利用(Utilization of evaluation):在日常事务管理层次和更宽泛的筹资或政策层面,决策者以及其他项目各方对于评估概念、方法和具体评估结果的使用。

准备评估 2

每项评估都要适合项目本身的情境。评估者承担任务的多少取决于评估的目的、项目的概念化和组织化结构,以及可以利用的资源。所以,评估者首先要与评估主办方和政策制定者、项目人员以及项目参与者等其他项目方探讨评估环境的各个方面,阐明评估方案。通过与主要项目方探讨和协商,评估者就能制订方案、说明评估将要解决的问题、明确解决问题的方法以及评估过程中和项目各方的关系。

就评估设计而言,还没有一种万能的"模板"。无论如何,只有关注特定的关键问题,才能使评估方案和项目情况一致。例如,评估方案要体现评估目的,关键还在于让评估主办方和其他重要项目方对评估有所理解。评估的目的在于为项目决策者改进项目实施提供信息反馈,而不单是帮助投资人决定终止哪个项目。另外,评估规划还要反映项目设计和组织思路,这样,提出的问题和资料搜集工作的安排才会与项目情形相适应。最后,理所当然,任何评估的设计都会受到时间、人员、资金和其他类似资源的限制。

尽管在细节上变化多样,但评估者在评估中会遇到相似的项目环境。因此,在准备评估阶段所得到的评估设计只是一系列类似评估步骤或主题中的一种或一部分。实际上,对评估的准备(调试性评估)基本上是一个针对被评估项目特定环境选择和进行评估设计的过程。一方面,主要是围绕评估者与项目各方的关系状况来选定评估方法;另一方面,要围绕评估问题的一般性组合和解决这些问题的通用方法来组织评估。本章概要描述的正是评估者在准备评估时需要考虑的主题和因素。

对评估而言，最具挑战性的一点就是，评估没有“通用”方法。每项评估的环境都有其独一无二的特点。因此，评估设计必须考虑到具体评估情景与评估者的方法、技术和概念全面运用之间的相互影响。好的评估设计既能适合评估环境，又能找到解决问题的可信和有效方法，最终达到改进项目的目的。与这个目的相关，本章讨论了在准备评估时需要考虑的主题和因素，描述了评估者的评估方案所应该包括的内容和评估的社会背景。

评估方案必须包括哪些方面

评估设计可以很简单，譬如，是否要使用计算机辅助教育项目来帮助三年级学生改善阅读。评估设计也可以相当复杂，例如对在多个城市聚居区开展减少药物滥用的一系列项目活动和效果的全面评估。不过，从根本上说，任何评估设计都包括以下三个内容：

评估要解决的问题。大量有关社会项目的问题都由各政党提出。所涉及的问题很多，包括项目**对象**(Target)(比如个人、家庭或社会群体)的需求以及他们是否被充分顾及和服务、项目的管理和操作、服务的效果如何、项目是否达到预期效果、项目的成本和效率，等等。没有一项评估能够或者应该囊括以上各个方面。评估涉及的关键一点就是：评估的基本目的和相关问题的细化，以及所关注问题的优先次序。

评估中用来解决问题的方法和程序。评估者的一项重要专业技能就是，知道如何掌握被评估项目各方面表现的有效的、及时的、可信的信息。社会研究技术和概念化工具的全面运用，有助于这一任务的完成。评估设计必须选择适当的方法来解决问题，并在工作规划中进行详细描述和组织。而且，所选方法不仅要切实可行、能够帮助解决有意义的问题，还必须在准确反映评估情况的前提下，使评估本身具有一定的科学严谨性。

评估者与项目各方关系的性质。在几十年的评估经历中，最重要的教训之一就是，尽管评估者假定项目各方会对评估发现感兴趣，但项目各方对评估结果的认同和应用从来都不是自觉的。因此，评估设计的一部分内容，就是规划与项目各方的有效互动，通过互动来提出和归类问题、实施评估、有效地利用评估结果。这种互动可以是高度合作性的：对项目各方来说，评估者只是扮演了顾问或推动者的角色项目方要对计划、实施、利用评估承担首要责任。评估者也可以承担上述职责，但需要从项目各方那里获得重要意见和信息。另外，评估规划还应该指出评估结果的发布程序(包括什么样的受众、在什么时间、接受什么信息)、书面报告和口头简报的内容及时间表，以及除了评估主办方之外，可以在多大范围内公布评估结果等内容。

评估方案应该考虑哪些问题

评估方案设计必须在仔细考察评估的社会背景基础上进行,要制订评估规划,就要对评估活动进行仔细分析。在分析中,需要考虑的重要因素有三类,分别为:①评估目的;②项目结构和条件;③评估可用资源。

评估目的

发起评估有很多种原因,譬如帮助改善项目管理、为支持者或批评者辩护、获取项目效果信息、提供决策所需的项目基金、提供关于项目结构或行政管理的信息、回应政治压力。因此,评估者需要说明的首要问题之一,就是评估目的到底是什么。在对这个问题的回答上,有时候简单,有时候却很复杂。对评估目的的陈述通常包括了对评估原因的说明,但这种书面表述的目的并不是整个评估的真正目的,有时甚至只是修辞性的。此外,在某些项目情境下,评估就是日常工作的一部分,没有任何特定理由,也与主办方的意愿无关(参见专栏 2—A)。

专栏 2—A　有人需要这样的评估吗

通过与社区服务部官员的最初会见,我们了解了进行评估的大概原因。他们说,需要的不仅是项目信息,而且是每个项目的实施以及成本—收益的信息……逐渐清晰的是,最关注评估的人是监管部门主管合同的官员,但我们并不知道到底怎么评估,从哪里获得评估需要的具体信息。我们只是认识到,政府资助的评估是被授权的,但并不知道授权者是谁。

资料来源:Dennis J. Palumbo and Michael A. Hallett, "Conflict Versus Consensus Models in Policy Evaluation and Implementation," *Evaluation and Program Planning*, 1993, 16(1): 11-23.

对评估目的的确定,首先要尽量知道谁需要评估、需要什么样的评估以及为什么需要等问题。尽管回答这些问题没有固定的模式,但一般而言,像记者那样从发掘故事开始,往往是最好的途径。也就是说,查找文献资料、访问主要知情人、挖掘客观的历史和背景资料。虽然在细节方面会有很大出入,但评估的原因无外乎以下一个或几个主要原因(Chelimsky, 1997):项目改进、责任承担力、知识生产、隐秘的动机。但是有时候,其他一些相当不同的动机也会起作用。

项目改进

评估的结果要为项目改进提供指导性信息。如果评估的主要目的就是帮助项目运作得更好,那么这样的评估就是塑造性评估(Scriven,1991;参见专栏 2—B)。塑造性评估结果所针对的受众,主要是项目策划者(在项目策划阶段)或项目管理者、监察委员会或对优化项目绩效感兴趣的投资者。这些人所提供的信息与对项目的需求程度、项目概念和设计、操作化、结果或效率等相关。在

这种情况下,评估者通常要和项目管理者以及其他项目方密切合作,共同设计、实施、汇报评估活动。致力于项目改进的评估,通常注重结果的及时性、具体性、即时应用性。所以,在评估过程中,评估者和代表性受众在结果上的互动会很频繁,且相对非正式一些。

专栏2—B 无人拨打的戒烟热线

通过塑造性评估来帮助设计一条"戒烟"热线。这个项目由健康维持组织(HMO)提出,目的是帮助癌症控制项目中的2 148名吸烟者。通过专题小组方式,确定了热线顾问的电话讲稿和其他服务,并用电话访谈的方式对HMO吸烟人员的代表性样本进行了访问。根据受访者的反馈,又进一步修改了讲话稿(更精炼),并根据参与者的建议,尽可能地对时间进行了重新安排;随后,通过实时通讯和其他方式,将所要提供的服务通知所有项目参与者。即使如此,热线开通的33个月中,平均每月也只有3个电话,目标人群的使用率仅为2.4%。为了更深入地评估这种令人失望的结果,在相同的范围内,用类似的服务进行了对比研究。结果发现低使用率是很典型的,但另一种热线涉及更多的人群,所以接到了更多的电话。项目主办方由此认为:为了使项目成功,吸烟者热线本就应该包括更多的人群,并且应该进一步引起公众的注意。

资料来源:Russell E. Glasgow, H. Landow, J. Hollis, S. G. McRae, and P. A. La Chance, "A Stop-Smoking Telephone Help Line That Nobody Called," *American Journal of Public Health*, February 1993, 83(2): 252-253.

责任承担力

社会项目虽然花费了纳税人的钱,但同时,它们也会为社会带来收益。因此,项目经理就希望有效地利用资源,并让资源产生实际收益。有一类评估就是为了评价以上预期是否达成,由于其目的在于对项目绩效做出综合性评价,所以,这类评估通常被称为总结性评估(Scriven,1991;专栏2—C就是一例)。总结性评估的结果通常提供给决策者以及对项目进行监督的主体(例如基金机构、管理委员会、法律委员会、政策决策者或更高的管理层),但也可以是正式决策制订圈外有兴趣的批评者、选举人和相关市民。评估的结果会对重要决策构成直接影响,例如项目的继续、资源分配、重组或法律行为。因此,这类评估需要相当充足并符合科学标准的可靠信息,以便提供项目评估的合理性基础,且以此来反击对项目合理性的质疑。评估者要相对独立地制订、实施评估以及汇报评估结果,但不会直接参与项目各方的决策活动。简言之,避免不成熟的、粗糙的评估结论是很重要的。因此,在给代表性受众反馈评估发现时,就应该通过正式的渠道、用书面报告的形式,而且要等到评估结束以后。

专栏2—C 美国审计总署对乳房X线照相质量标准的早期效果评价

1992年,乳房X线照相质量标准机构要求食品与药品管理局(FDA)在美国所有州为乳房X线照相程序制定单一标准的编码。法案通过后,国会不得不重新关注这个问题,而且乳房X线照相服

务的市场销售下降了,因为供应商宁可降价也不愿意改善操作以适应新标准。美国审计总署应邀评价执行法案的初期影响,并向国会报告。他们发现,FDA 在采取渐进方式执行法案的要求,以帮助减少因销售下降所带来的不利影响。FDA 检察员没有关闭不符合标准的设施;而宁愿给他们更多的时间,去改正已经发现的问题,以符合新的质量要求。只有很少机构停止了他们的乳房 X 线照相服务,而且通常都是小型供应商,且 25 英里以内还有其他持合格证的机构。GAO 的结论是,在乳房 X 线照相服务质量上,乳房 X 线照相质量标准取得了积极的效果,和国会所期望的一样。

资料来源:U. S. General Accounting Office, *Mammography Services: Initial Impact of New Federal Law Has Been Positive*. Report 10/27/95, GAO/HEHS-96-17 (Washington, DC: General Accounting Office, 1995).

知识生产

有些评估的目的不是直接为决策及其相关方面提供信息,而是因为更复杂的目的和向更多的受众描绘社会干预的性质和效果,从而增进相关知识。举例来说,一位学者可能根据理论,对一个创新性的科学课程进行检测和检验(参见专栏 2—D 的一个例子),判断项目是否能起到作用,效率如何?当学者带着自己的兴趣来研究这样的社会干预(如一门创新性的科学课程)时,同样的情况发生了,学者想要的是对知识增长有所贡献。类似地,政府机构或私人基金也会针对某一社会问题设立并评估示范性项目,从而研究新的解决方法。因为这类评估要对社会科学知识做出贡献,所以通常要在科学的框架中,运用可行的、最严格的方法实施。如果是示范项目或外界资助的研究,研究结果的读者就包括研究主办方、学者和政策制定者。在这种情况下,评估结果的发布最可能通过学术期刊、研讨会以及其他类似的专业渠道。

专栏 2—D 检验针对病态性赌博的创新性治疗观念

造成病态性赌博的主要原因包括:赌博的刺激、无所事事、家庭和工作打击、偷钱以及想赢回本钱。尽管近来赌博人数的增加已经导致病态赌博的上升,但是尚没有实际的治疗项目来帮助这种无序状况的受害人。针对赌博的心理学研究已经表明,问题赌徒有一种幻想,就是他们能找到提高胜算的策略,而不管游戏机所固有的随机性作用。据加拿大一个临床研究小组推测,建立在对错误观念进行"认知修正"基础上的疗法很可能奏效。因为过分的赌博会导致经济问题和人际关系困难,所以项目干预将针对赌徒的认知干预活动和问题解决、社会技能培训联合进行。

为了检验治疗的概念,研究者利用媒体广告和健康人的介绍,招募了 40 个愿意接受治疗的病态赌徒。治疗组或对照组是随机指派的,在治疗之前和之后,按照不同的时间间隔,调查了病态赌博者的赌博方法、对控制的理解、赌博的期望、对自我绩效的理解以及赌博频率。结果显示,治疗组在 6~12 个月时间内,取得了积极的改变。然而,与预期效果相比较,结果打了很大的折扣。在治疗组中,20 个赌徒只有 8 个戒掉了赌瘾;而在对照组中,20 个成员中有 3 个放弃病态赌博。这样的情况在对成瘾性问题进行社会干预时,会经常遇到。尽管如此,研究者认为,所获得的结果足以证明他们的治疗概念是有效的。

资料来源:Caroline Sylvain, Robert Ladouceur, and Jean-Marie Boisvert, "Cognitive and Behavioral

Treatment of Pathological Gambling: A Controlled Study," *Journal of Consulting and Clinical Psychology*, 1997, 65(5): 727-732.

隐秘的动机

至少对那些评估主办方来说,有时候,评估的真正目的和项目实施情况的信息获得过程无关。例如,项目管理者或委员会要发起评估是很正常的,因为他们认为这是好的公共关系策略,并且能给投资方或政策决策者留下深刻印象。偶尔,有的评估也只是为早已存在的幕后决策(如终止项目、解雇某人)提供公共环境。还有的情况是,把评估看做是用来安抚反对者和推迟不同决策的拖延策略(既不任命一个委员会来研究问题,也不针对问题采取行动)。

事实上,在所有发起评估的动机中,都有政治操作和公共关系的因素。但是,如果这些因素成为首要目的,评估者就陷入了两难境地。评估既要由政治或公共关系目标指导(这就可能使完整性受到影响),又要关注项目实施中的问题(发起评估、且不关心评估的人甚至会威胁到评估)。无论是哪种情形,评估者都要尽量避免陷进去。如果在评估探索阶段就出现重要目标缺失的状况,继续制订评估方案通常是不明智的,有远见的评估者就可能会降低参与度,甚至终止参与评估。另外,评估者也可以假定承担咨询者或顾问的角色并且帮助相关方面澄清评估的实质、明确区分现实的和理想的期望,并努力调整,使评估起到更适当的作用。

项目的结构和环境

即使表面上都提供了"相同的"服务,也没有两个项目的组织结构和环境、社会政治条件是相同的。项目结构和环境的细节,构成了评估环境的主要特点,因此,评估方案据此进行调试。虽然这样的细节很多,但对评估者来说,从对评估设计和执行的普遍影响来看,尤其重要的是三大类:项目进展的阶段,项目的行政和政治环境,以及项目的概念和组织结构。

项目进展的阶段

社会项目过程可视为阶段性进展的累积,不同的阶段提出不同的问题,因此,要解决这些问题需要不同的评估方案(参见专栏 2—E)。对处于初期计划阶段项目的评估就截然不同于对既有项目的评估。同样,对重组项目的评估也不同于对运作和功能基本稳定的评估,二者的关注点会完全不同。

在新项目,尤其是开创性项目的开始阶段,评估者通常要检验项目所能满足的社会需求、项目方案及目标、目标人群的定义、预期产出以及获得这些产出的方式。在这一阶段,评估者在项目实施之前或者项目实施之初,可以通过评估和改进项目方案,扮演项目咨询顾问的角色。

专栏2—E 项目发展阶段和相关评估的功能

项目发展阶段	问　题	评估的功用
1. 社会问题和需求评估	在多大程度上满足社区需求和标准?	需求评估、问题描述
2. 确定目标	怎样才能够满足那些需求和标准?	需求评估、服务需求
3. 项目备选方案	用什么样的服务来产生预期的变化?	项目逻辑或理论
4. 方案选择	什么样的方案可能是最好的方案?	可行性研究、塑造性评估
5. 项目实施	如何实施项目?	实施评估
6. 项目运作	项目是否按计划在运作?	过程评估、项目督导
7. 项目产出	项目是否获得了预期的结果	产出评估
8. 项目效率	项目的结果与成本比较是否合理?	成本收益分析、成本绩效分析

资料来源:S. Mark Pancer and Anne Westhues, "A Developmental Stage Approach to Program Planning and Evaluation," *Evaluation Review*, 1989, 13(1): 56-77.

有时候,新项目的评估是要回答结果和效益问题,但项目早期的不确定性通常会使得对这些问题的回答不成熟。对新项目来说,正常的情况是,要用一年或更多的时间建立设施、招聘和培训员工、与目标人群订立合同以及使服务达到预期水平。在这期间,项目对社会条件方面的改善可能没有预期的那么理想。过程评估致力于认清目标人群的需求、改进项目操作以及提高服务质量,在第4—6章所讨论的方法就更适合这些案例。

虽然新项目评估是一种重要的评估活动,但到目前为止,对既有项目的评估仍然占据着更重要的位置。评估这样的项目,首先要了解它们的社会和政治历史。大多数成熟的社会项目都经历了长时期的改善,除非发生的一些危机或者必须考虑做根本性的改变,否则就会严格遵循传统方式和方法推进。通常来说,如果有人对部分项目方的基本假设和活动方法有任何质疑,都会受到激烈的反驳。不过,像社会保险金、学校指导顾问、残障人士就业项目、劳改犯的假释监督以及预防疾病的社区健康教育之类的保障项目,其价值是公认的。但是,如果项目庞大且成熟,评价其效果和效率就会比较困难,尤其是向所有项目对象提供服务的全面覆盖项目。在这种情况下,项目往往缺乏基本的标准,这就使得评估者准确叙述项目产出时受到很大限制。通常而言,这类项目的评估会指向项目目标的明确程度,与项目主办方、员工、其他项目方兴趣的相关程度,项目实践和计划是否一致,以及项目是否送达到目标人群等。例如,美国农业部对食物券项目进行阶段性的参与研究,以测量目标家庭的参与程度,从而扩展项目服务,提高参与度(Trippe, 1995)。

然而有时候,由于政治攻击、竞争、准备项目经费或来自目标人群意想不到的变化等外在压力,在确定项目时,就要进行评估,因为项目状况已经成为问题。

或者,因为项目主办方和成员对社会干预的绩效不满,希望作出改进。无论是哪种情况,都要考虑进行部分调整,通过评估来指导这种变化。在以上各种情形下,评估可以将重点放在项目的各个方面:项目是否必要、项目的概念化和设计、项目的操作和实施,或者项目的影响和绩效等。

例如,上面提及的联邦食物券项目就是一项实行了二十几年的国民项目,目的就是通过给贫困家庭提供可兑换的食物券,让他们在杂货店购买合格食品,进而提高贫困家庭食物消费的数量和质量。农业部打算放弃食物券而代之以支票,以便减少印刷、分发和兑换货币的较高成本。为了检验这样做的效果,前后共进行了4次实验来对比接受食品券的家庭和接受等值支票家庭的食物消费(Fraker,Mratini,and Ohls,1995)。结果表明,两者之间存在明显差别:接受支票的家庭比接受食物券的家庭购买的食物少。农业部因此决定继续保留食物券,尽管推进这个项目的成本更高。

项目的行政和政治环境

除了学者能够按照知识生产的初始目的实施评估研究以外,其他评估者在建立有关项目是什么、项目的目标和评估问题是什么这些内容的定义时,是不自由的。因此,评估者必须和评估主办方、项目成员以及其他项目方互动,共同建构评估的基本背景。在大多数情况下,不同群体的观点总是略有差异,评估者要力图制订一个反映所有重要观点和利益关系的评估方案,至少要和主要项目方的主流观点保持一致。

如果主要项目方在任务、目标、程序或项目问题等关键方面存在本质的冲突,那么,评估设计就会极度困难(参见专栏2—F)。尽管评估者可以尽力将冲突的观点整合到规划中,但很不容易,因为,评估主办方并不希望接受来自敌对群体的问题和观点。何况,在很大程度上,几乎不可能将这些不同的问题和观点整合到一个评估方案中,即使可能,也需要更多的时间和资源。

专栏2—F 项目方在家庭监管项目上的冲突

在对假释犯人使用电子跟踪器的家庭监管项目评估中,评估者对项目各方的观点提出了以下意见:

大量相互冲突的目标(包括降低成本和监狱转移,控制和公共安全,中间惩罚和增加的改造选择权,治疗和康复)对不同的机构来说具有不同的重要性,不同的项目方强调不同的目标。有些立法者强调降低成本,另一些人则强调公共安全,还有一部分人重点关心将犯人从监狱转移出来。有些执行者强调对控制的需求和对某些"功能紊乱"个体的约束,同时,另一些人则关心康复与犯人的再社会化。因此,让"主要政策制定者、管理者和项目职员"在优先问题和改进项目的途径等问题上达成一致,基本不可能实现。

资料来源:Dennis J. Palumbo and Michael A. Hallett, "Conflict Versus Consensus Models in Policy Evaluation and Implementation," *Evaluation and Program Planning*, 1993, 16(1): 11-23.

评估者也可以从项目某一方的观点出发设计评估,尤其是参照评估主办方

或评估主办方指定的项目方的观点。这样，当然就不会得到持对立观点的其他项目方的大力支持，甚至他们会反对评估、批评评估者。对评估者来说，不管评估的目的是什么，由于批评直接指向评估所代表的观点和评估的理由，所以，挑战是非常明显的。评估主办方坚持从自己的观点出发设计评估方案，并不必然是不适当的；同样，评估者从项目某一方的观点出发设计和实施评估，而并不太顾虑反对意见，也并不必然是错误的。

例如，假设为长期失业人员提供职业培训项目的主办方所关心的是，项目是否只是挑选了一些容易对付的个案、提供的服务是否主要是职业咨询而不是技能培训，以及组织是否无效率，主办方会非常及时地委托一项评估来检验这些问题。另一方面，项目人员和他们的支持者可能具有极端对立的观点，来证明个案的选择、训练方式和管理实施是正当的。在这种情况下，一个尽职的评估者应该听取管理者的观点，并鼓励他们参与评估，这样，就可以使得评估方案尽可能贴近项目本身。因此，项目内容和操作的合法性，一定会受到管理层的关注。但无论怎样，评估方案的基础是评估主办方的观点及其关注的问题。评估者的首要责任仅仅是明确评估所采用的观点，不让误解产生，并对项目成员公正和诚实。

对评估者来说，处理项目各方冲突状况的另一种方法，就是尽量提供一种可能性，让冲突各方就有争议的部分能够相互理解。这是可以做到的，譬如说，努力阐明各方不同的利害关系、假设和观点。在这样的评估中，评估者通常要检验并对所有群体通报评估的具体工作。例如，接受特殊教育儿童的父母认为，他们的孩子受到了主流的、正常班级的诬蔑和歧视。教师会同样强烈地认为不是这样。实地观察的结果表明，正常儿童和接受特殊教育儿童的互动的确存在问题，但问题仅仅发生在教室之外的操场上，而且是非正式的互动。

如果项目各方的冲突根深蒂固且是敌对性的，则这样的冲突有可能来自政治价值或意识形态上的深刻差异。也就是说，无论问题怎样广泛和普通，评估工作也无法将其整合起来。有一派评估者认为，所有的评估环境都是这样的，这也是评估者必须面对的基本特点。从这一角度来看，项目指向的社会问题、项目本身以及项目的意义和重要性，都有明显来源于不同个体和群体的社会意义。因而，评估者与其将注意力放在项目的目标、决策、结果诸如此类的方面，不如直接关注不同项目方提出的意见、观点、问题和价值。

作为评估事业这一特殊事物的提倡者，古巴和林肯（Guba and Lincoln, 1987, 1989, 1994）的研究认为，评估者的正确职能是推进项目各方之间的解释性对话。因此，尽管项目各方仍然保留着自己的意识形态和观点，评估的基本目的就是要促进项目各方之间的沟通，借此获得项目价值和意义的共识。

最后，评估者必须认识到，尽管项目各方会尽最大努力来沟通和制订适宜的评估规划，但各方始终会坚持自己最基本的观点和政治立场。这就意味着，如果结果与评估主办方和其他项目方提倡的政策及观点发生冲突，他们会将矛头指向评估者，并严厉地批评评估活动。因此，即使评估者在和项目各方共事，并很好地把他们的观点纳入了评估规划，在结果出来的时候，他们也别期望得到英雄

似的赞扬。由于项目各方观点的多样性,无论出现什么样的结果都会导致部分人的不满。评估者在政治环境中工作,这是这类环境的实质,项目各方通常会反击与他们利益相反的证据,强烈地质疑证据和提出证据的那些人。即使每个人事先都同意评估问题及其解决方案,或者每个项目方都明白真实的结果可能不会支持他们的立场,也起不了多大作用。即使如此,评估者也应该从开始就辨明项目各方的立场,制订策略来减少评估过程中由于不同观点所带来的意见不合,限定他们对评估结果的期望。

项目的概念结构和组织结构

显而易见的是,如果项目方对项目活动没有一个明确的概念,那么就很难评估项目开展情况的好坏。所以,评估方案设计的要素之一就是项目的概念化或项目理论,亦即项目方案的操作化和明确化、指导项目行动达成预期结果的逻辑以及项目活动整体安排。这种概念化结构或**项目理论**本身,就是评估的焦点。项目概念结构越清晰,越令人信服,评估者越容易识别项目功能和评估应该关注的项目效果。项目概念化是否适合项目所指向的社会问题?如果对这个回答有明显的不确定性,那么在评估设计中关注项目实施的进展等问题就没有什么意义。在这种情况下,不如将评估活动转向项目规划本身,看看如何规划才能更有效。在新项目的计划阶段,评估者的参与通常有利于概念化的形成,进而使得在确定项目实施等主要问题方面不但更加明确,而且更加有效。

在计划阶段之后,尤其是对成熟的项目来说,项目成员或主办方一般不会再去系统地纠缠基本假设和期望,而是把主要的注意力便转向日常行为和常规操作程序,项目成员很难修改既有的项目原理,也很难只同意其中的任意某个部分。例如,根据学校和咨询机构的合同,工作人员要应对有问题的儿童,他们能做到很清楚地向当事人讲述他们的咨询理论、目标和治疗技术。但是,他们很难提出和赞同这样的观点:如何进一步改善家庭沟通状况,进而获得更好的结果。随后,帮助项目工作人员形成完成项目活动所需的隐含的基本原理,就成为评估的任务。

设计评估的时候,还必须从具体的角度考虑项目的组织结构。就项目问题性质的了解和覆盖面、资料搜集过程、评估所需资料和涉及的项目各方来说,多元化的服务或目标人群、分散的服务场所或设施以及和其他组织实体广泛的项目合作等这样的项目特点,都有很强的参考价值。更大、更复杂、更集中、在地理上更分散的组织结构,相对于简单的情景来说,呈现出更多的实践困难。在这样的情况下,评估工作通常需要一组评估人员以及与项目大小和复杂程度相适应的资源和时间。在评估文化中,复杂和多维度的项目给评估带来的挑战就在于:项目的多个维度明显分属于不同的评估研究领域(参见专栏 2—G;Turpin and Sinacore, 1991)。

专栏2—6 刑事审判的多角度评估:成功的结构化障碍

在实施可信的评估中,除了考虑基本方法以外,如果对罪犯进行法律制裁被强加上社会、政治和组织的限制,就会使得多角度的评估很难、很冒险。一开始,系统就呈现极度多样化的状态。譬如,警察局要和自治州、县、大学校园、公共房屋、主要运输线以及联邦政府打交道。法律制裁体系又高度分化:城市管理警察局和监狱;州政府管治安和检察办公室、监狱和缓刑办;联邦政府管理整个监狱系统,并在不同的政治机构安插办事处,每一个办事处在缴税和开销上又有自己的规定。而且,由于法律制裁办对他们的工作奉行保密的亚文化,就会对评估活动的影响很大,因为评估者被视为管理层、法院或抱有政治目的的个体派来的"私家侦探"。站在一条线上的人很自然地就采取"我们反对他们"的态度来排斥外来的评估者。法律制裁办通常也处于高度命令化的政治环境中。他们是当地政府最可见的组成部分,而且花费最多,他们的行为总是受到历来被视为"看门狗"或反对派媒体的监视。最后,法律制裁体系在这样的背景下产生作用:在程序问题上存在个体权力—法律制约的相互作用,对个体来说不愿意以身犯险,而法律制裁体系则有义务提供个性化处理(虽然不是实际上的)。这就会将自由或公正概念的实践引入一种让人厌恶的状况:项目影响评估的最佳方案不得不屈从于项目效果方面可解释的信息。

资料来源:Wesley G. Skogan and Arthur J. Lurigio, *Multisite Evaluations in Criminal Justice Settings: Structural Obstacles to Success*, New Directions for Evaluation, no. 50 (San Francisco: Jossey-Bass, summer 1991), pp. 83-96.

项目提供的具体干预以及服务的性质与结构也同样重要。对评估来说,最简单的干预是那些间断的、一次性的事件(例如为无家可归者提供食物),可以预期产生相对迅速的可见效果(他们不饿了)。通常,这类干预的组织行为和送达系统相对直接(救济贫民的施舍处),服务本身也不复杂(提供食物),而且产出明显(有人吃了)。这些特点简化了评估要提出的问题、资料搜集工作和结果解释的复杂性。

对评估者来说,最难以评估的是那些分散的(社区组织)、跨度很长的(小学数学分数)、应用很广的(折衷主义精神疗法)、取得预期产出需要很长时间的(学前补充教育)、不明显的(生活质量的提高)项目干预产出。对这类干预来说,由于服务本身的不确定性及其潜在效果,涉及项目过程和产出的评估问题就很多。而且,由于项目实施和产出的复杂性或分散性,评估者很难找到清楚明白的、关键的评估测量方法。如果要在规定的时间内搜集包括大量不同变量和观测值的资料,也极具挑战性。所有这些因素都会影响到评估方案,尤其影响到完成评估方案的过程、难易程度和所要调用的资源。

可用于评估的资源

要实施评估必须有一定的资源。无论是从项目现有资源中或评估主办方或者独立基金中,评估者必须获得大量用于评估活动的资源,包括时间和参考资料、装备和实践资料的搜集、分析和汇报的工具。因此,评估规划的重要方面除了任务和时间安排以外,还要按照计划实施的步骤,对人员、资料和开支做出详

细的预算。当然，所需资源的总和应该在可获取的范围内，否则就要对计划或资源配置作适当的调整。赫德理克、比克曼和罗格（Hedrick，Bickman and Rog，1992）、卡尔德、格林和彼得森（Card，Green and Peterson，1992）以及芬克（Fink，1995，第9章），都对资源、预算和期限的确定等操作化议题提出了可以借鉴的意见。

当然，尽管经费是评估必须涉及的关键资源之一，但并不是评估者关心的唯一资源，认识到这一点是很重要的。评估是一种特殊的调查方式，只能发生在被评估项目的可操作环境中。这就意味着，举例来说，要想做好评估，就要掌握有关的技术和经验。在大型评估项目中，大量的专业评估者、资料搜集员、资料管理者、分析师和助手都需要完成高质量的工作。即使有充足的经费，要找到足够的专业人员，也并不总是一件容易的事情。这就是通常要把大型复杂评估项目交给专门的研究机构去做的原因，因为他们掌握着大量的专业人员。

评估的另一个关键资源就是项目管理层、职员和其他项目方的支持。举例来说，项目成员在搜集资料方面（如观察主要项目活动的机会）的合作程度，就能对评估能完成到什么程度有很深的影响。虽然这些因素不像货币价值那样容易表达，但对评估来说，的确是有价值的资源。获取资源的障碍和项目合作的缺乏或者更有甚者——主动的对抗，都会使评估活动付出相当大的代价。要克服这些障碍、有效地完成评估，就需要花费相当多的时间和努力，更不用说对所有因素的关心所带来的压力了。在更极端的例子中，这类对抗会危及评估的适用范围甚至合法性，使得评估者无法完成评估。

评估者与项目方之间一个尤其重要的互动就是获取和使用项目记录、文件及其他内部资料。这样的资料对于确定服务对象的数量和特点、接受服务的类型和数量以及提供服务的成本来说，都是必不可少的。如果能够从项目记录中获得可靠信息，就不需要评估者再组织单独的、花费更多的资料搜集工作。使用项目记录的难易程度不同，手写记录如果不经过大量的编码处理，就很难应用。相反地，可读的数据库存储的记录是机读资料库，比较容易处理，不需要太多的数据整理。值得做的是：考察项目的真实记录，判断记录的完整性和可用性，进而确定评估的过程与合作的方式。

最重要的一点是，评估者在设计评估时，必须将项目成员的配合、项目资料的获取以及从项目记录中所得资料的性质、质量和有效性，看做是重要的资源问题。如果从一开始就和评估主办方、项目成员及其他项目各方讲清楚评估各方面所需的资源和支持，就会把误解和对抗的可能性降到最低（Hatry，1994）。

除了充足的经费和项目成员的合作之外，有经验的评估者知道，还有一个最珍贵的资源就是时间。在制订评估方案时，对完成评估的时间段和机动性要有基本的考虑。不过，这些事情几乎不由评估者的意愿来决定。与评估有关的决策要服从政策议事日程的需要。因此，就不得不按照某些决策者的意思，在特定时间，让评估结果在决策中扮演角色；在这之后，这些结果也就没有用武之地了。这种约束通常体现为实施评估的严格时间限制。更复杂的情形是，评估主办方

和决策者都低估了完成评估需要的时间。极为平常的是,评估主办方要求在几个月内完成大量问题的探讨、付出相当多的努力并获得评估结果。

公平交易十分重要。评估能做到有广度、深度和严密性,但需要与之成比例的时间和资金。或许评估能很便宜、也能迅速完成,但那只是应付很小范围的或者相当肤浅(或兼而有之)的问题。除了最有经验的评估主办方之外,几乎所有的评估主办方都希望评估既要有深度、广度和严密性,还要便宜、快速。正因为如此,结果常常是,在资源不足的情况下,为追求进度,评估者不得不发疯似地超负荷工作;在接受他们花钱所买的评估结果时,评估主办方不得不因为结果的不足和拖延而感到焦虑。在评估可用时间和专业技术以及实际计划的方法和程序之间,存在着特别直接的关系。无一例外,评估结果要达到的科学标准越高,对时间、技术、努力程度和项目合作的要求就越多。

所以,与其解决大量一般性的问题,不如解决有限的重点问题。因为评估主办方和其他项目方通常对实施高质量评估所需的努力和技术没有实际的认识,所以很容易产生误会和冲突。防止出现这种情形的最好方法,就是和评估主办方详细地商讨评估所必需的资源;或者在有限资源的条件下,进行公平交易。

评估者与项目各方关系的性质

无疑,每一个项目都是一个社会结构,其中,各种各样的个体和群体都致力于构成项目的角色和活动:项目管理者进行管理,职员提供服务,参与者接受服务,诸如此类。而且,每个项目也是一系列政治和社会关系组合,牵涉到与项目有关联或关注项目的主体(例如相关政策的制定者、竞争性项目和鼓吹群体)。在项目评估的早期计划阶段,评估者应该谨慎地把握与项目各方关系的属性,分清哪些人直接参与评估,哪些人仅仅对项目评估的进度和结果感兴趣。更具体地讲,评估必须考虑以下所列出的这些项目各方:

- 政策制定者和决策者:决定项目是否开始、继续、中止、扩展、调整或缩减的人。
- 项目主办方:发起和资助项目的组织。他们也可以是政策制定者与决策者的重合。
- 评估主办方:发起和资助评估的组织(有时评估主办方就是项目主办方)。
- 目标群体:被评估项目干预和服务的个体、家庭或其他单位。
- 项目管理者:监督和管理干预项目的项目成员。
- 项目职员:提供项目服务或扮演支持角色的项目成员。
- 项目竞争者:和项目竞争有用资源的组织或群体。例如,一个提供可选择学校的教育项目会引起公立学校的注意,因为他们把新学校看做是竞争对手。
- 项目各方:对项目在做什么或发生了什么事情有兴趣的当时环境中的组织、群体、个体或其他社会单位。

- 评估和研究群体:攻读评估专业并获得技术质量、可信度和学术认证的评估专家,以及在有关项目领域内工作的其他研究者。

无论是哪种情况,评估都会涉及这些人中的部分甚至全部。但是,无论把关注评估的个体或群体按照什么标准分类,评估者都必须计划好用何种方式和他们互动,并且了解他们的观点,对于与主要项目方进行正确互动方式的考虑应该是评估规划的一部分(参见专栏 2—H 包含项目各方的意见之一)。

专栏2—H　在评估中项目各方的卷入:实践性建议

基于与学区职员共事的经验,一位评估者提出了以下建议,通过项目各方的参与来让他们支持评估的利用:

- 识别项目各方:从一开始,就识别具体的项目各方,他们是对项目最关注的人,并承担了极大的风险。
- 尽早吸收项目各方参与:一旦对项目各方的识别完成,就让他们参加到评估过程中来,因为许多涉及评估的关键性决策都要在前期产生。
- 不断让项目各方参与:主要项目方的参与应该是所有评估阶段的一部分;如果可能,安排定期的集体会议。
- 让项目各方积极参与:项目各方参与的关键问题是要积极;项目各方应该参与设计问题、帮助进行调查、参与最终的总结和报告,以及商讨方案的重要方面。
- 建立结构:在与项目各方观点相似的基础上,设计和使用概念框架,能够有效地帮助双方持续对话。这一框架应该突显当时的关键问题,以便项目各方能够分享关注点和观点、明确信息需求以及解释评估的结果。

资料来源:Robert A. Reineke, "Stakeholder Involvement in Evaluation: Suggestions for Practice," *Evaluation Practice*, 1991, 12(1): 39-44.

考虑与项目各方的关系,必然要从评估主办方开始。评估主办方通常要提供经费、决定怎样做、什么时间做、由谁来做。与评估主办方的关系有各种可能,而且关系的具体形式极大地依赖于评估主办方的意愿及其与评估者之间的沟通方式。最一般的情形是,主办方希望评估者扮演独立的、职业从业者的角色,尤其在开始的时候,评估者要接受主办方的指导,但要对计划、实施和汇报评估结果负全责。例如,政府部门和其他项目资助者经常通过公布建议需求(RFP)或应用需求(RFA)来委托评估,相应地,评估者会按照他们的要求表述自己的能力、设计、预算和时限。那么,评估主办方就可以从中选择评估者,建立契约性的安排,以协调工作。

当然,为了使评估者和主办方之间的工作关系更加协调,也需要其他条件。例如,主办方也许希望参加评估规划、评估实施和结果分析的全过程,这样,既可以对评估者的工作进行逐步地回应,又可以实际参与每一步的工作。这一关系形式的变型,主要针对的是内部评估者,因为他们本身就是被评估组织的一部分。在这种情况下,评估者在规划和实施评估的过程中,无论合作者是作为评估

单位的管理层、被评估项目本身组织中地位较高的某些人或是两者兼有，评估者一般都要和管理者密切合作。来自组织外部的评估者也可以仅仅作为评估顾问，帮助评估主办方设计和实施评估，但不在工作中扮演主要角色。

有些情况下，评估主办方会提出与评估者合作共事，但合作者不是主办方本身，而是项目的另一个群体。例如，资助社会项目的私人基金会经常希望评估以合作方式展开，希望评估者与相关项目方有密切的互动。这种方法的与众不同之处在于，项目服务的接受者在评估计划、设置优先权、搜集信息和解释干预结果方面扮演了主要的角色。

评估者与评估主办方及其与另外项目群体之间的关系，是评估背景和计划过程的中心。所以，在专业性评估中便产生了一些特定的、总体上也没有系统性的词汇表，用来描述各种各样的评估环境。以下是词汇表中评估者与项目方关系的主要形式：

独立评估（Independent evaluation）。评估者全权负责制订评估方案、实施评估以及发布评估结果。如果某个社会科学家为了知识积累（自己出资或利用研究基金）而承担一项社会项目评估，评估者就能发起并自由地指导评估研究。但是，更一般的情形却是，主办机构委托独立的评估者，它们只规定评估的目标和内容，其他则由评估者自由执行（从具体规划到实施完成）。在这样的情况之下，在形成评估的过程中，评估者通常要和某个范围内的项目方协商，给他们施加一定的影响。

参与性或合作性评估（Participatory or collaborative evaluation）。这类评估方式是按照团队项目组织的，包括评估者和项目群体中一个或多个代表组成小组（Greene，1988；Mark and Shotland，1985）。参与其中的项目方和评估者合作完成评估计划、实施和评估的过程，而评估者作用的弹性很大：从小组领导或顾问到只是被叫来提供资源的人。参与性评估的一个著名方式就是巴顿（Patton，1986，1997）的“利用导向的评估”。巴顿的方法强调和某些特定的个体密切合作，这些人利用评估结果来保证评估反映他们的需要，并产生他们能够且会实际应用的信息。

授权性评估（Epowerment evaluation）。一些评估者已经提出了一个关于评估者—项目方关系的观点，强调项目各方的主动、辩护和自我决定（Fetterman，Kaftarian and Wandersman，1996）。在授权评估中，评估者与项目方的关系是参与性或合作性的。另外，评估者的角色还包括了对参与其中的项目各方能力发展的咨询和帮助，表现在能够让他们自己实施评估，有效地利用评估结果来获得支持和改变，在一定意义上，对项目的影响因素进行控制。所以，评估过程不仅要产生信息或结果，而且要提高参与者的自我发展能力和政治势力。从这一点引申开来，授权性评估最适宜吸收在项目背景中没有权力的项目方，包括项目对象或广义的受益者。

参与或授权性评估的一个最大贡献就是回应独立评估的一些问题。在有些情况下，独立评估的优势不明显，或者因为更具参与性的评估的优势，而显得名

不副实。在授权评估中,评估主办方或其他项目方中的一个或多个群体的直接参与,能够确保评估结果解决他们的问题,对他们来说是有用的。而且,授权评估还能创造评估结果的所属权,增强评估结果的意义,减少造成对抗的潜在因素。正如授权理论家所指出的那样,当不拥有正式权力的项目各方能够实施并使用评估发现的时候,通过增强他们的影响力和绩效,就能改变项目背景中的权力平衡。所以,很自然,评估主办方和评估者会具体考虑如何分配项目责任,安排评估者与项目方互动,有组织地进行评估。

是由独立评估者,还是由项目各方规划并实施评估,都会对决策、评估者的角色以及评估的关注点和特点产生极大影响。无论如何,评估所获得的结论,应该体现公认的评估概念和方法在具体项目中的应用。因此,我们将和项目各方共事的过程(无论作为独立评估者、合作者、帮助者或者资源提供者)与由这一过程产生的评估方案区分开来。尽管评估方案可能在这个过程早期就已设计并确定下来,或者随着项目过程的发展和进步而失去意义,但是对评估环境和项目来说,好方案的特点就是方案的制订和实施可以在过程中分离开来。在本章的余下部分,我们会论及一般性的规划,并涉及评估者的角色(假定是独立评估者或合作小组)。

除了项目各方之间的关系,还必须强调评估者与项目各方互动的另一方面:评估结果的传播和发布。为了让评估结果产生作用,就必须把评估结果传播给那些关注项目的人,尤其要传播给那些项目的重要决策者。想要很详细地传播评估结果是很困难的,何况在评估完成的时候,哪个项目方对什么样的信息感兴趣,始终是不确定的。所以,最好的办法是和主要项目方讨论,并从开始就制订一个有组织地传播及发布评估结果的计划。关于筹划有效传播和发布活动的细节,可以参见托里斯、普莱斯基尔和皮翁德克的有关讨论(Torres, Preskill and Piontek, 1996;参见专栏2—I)。

专栏2—I　与项目各方的成功沟通

托里斯、普莱斯基尔和皮翁德克(Torres, Preskill and Piontek, 1996)调查和访问了美国评估协会成员,了解了他们和项目方沟通以及汇报评估结果的经历。明确了以下几个有效沟通的因素:

- 在评估进行中和合作性评估中进行沟通最成功。可以利用定期会议和非正式对话,在评估过程中维持密切联系,随着评估的进行,还可以利用中期备忘录和报告草案来传播结果。
- 运用多种沟通方式也很重要。可以包括简短报告和总结、口头汇报以及非正式的互动。
- 传播的内容应该适合受众,并且让他们很容易理解。要使用清晰的语言、图表以及生动而具体的阐述进行传播。能够提供项目和评估的背景信息,包含肯定的或否定的结果,做出有针对性的建议。

资料来源:Rosalie T. Torres, Hallie S. Preskill, and Mary E. Piontek, *Evaluation Strategies for Communicating and Reporting: Enhancing Learning in Organizations* (Thousand Oaks, CA: Sage, 1996), pp. 4-6.

评估问题和评估方法

项目评估其实是一个信息汇集和解释的过程。在这个过程中，要试图解决指定的、有关项目实施和效果的一系列问题。所以评估的一个重要步骤就是，确定评估必须解决的问题。有时候，对评估问题的确定非常草率，但我们提倡以认真和具体的态度来对待评估问题。经过详细建构的一系列**评估问题**(Evaluation questions)，会给出评估的结构，引导制订正确而周全的计划，并为人们对所关注问题的解决和解决方案的应用等问题的讨论，奠定基础。实际上，要使评估活动适宜于具体的评估环境并获得审查通过，组织评估问题并筹划如何回答这些问题，才是最基本途径。

对评估规划来说，由于明晰所要解决的问题是如此关键，第3章将首先全面讨论评估问题(怎样提出问题，以及怎样提炼、组织和整合问题)，然后才讨论评估方案。由于当前的目的不在于此，我们假设已经确定了一系列适当的评估问题，现在考虑这些问题的更广泛涵义，以调试评估、设计评估。对指定的项目来说，评估要解决的问题必然针对项目的特殊性质。评估中，一般概念和方法框架下，常见评估问题的类型如下：

- 需求评估：回答项目运作所需的社会条件以及项目需求程度等问题。
- 项目理论评估：回答项目的概念化和设计等问题。
- 项目过程评估(或过程评估)：回答项目的操作、实施以及服务送达等问题。
- 影响评估(产出评估)：回答项目产出和影响等问题。
- 效率评估：回答项目成本—收益和成本—绩效问题。

在本书后面的章节(第4—11章)中，我们会对这些评估问题进行详细讨论。在这里，我们只是提出每一类型评估最适合的环境。

需求评估

发起或维持社会评估的基本理由是为存在已久的或刚开始的社会问题(借此我们意识到社会某些方面所存在的不足)提供解决方案。例如，一个提高阅读能力项目的推动力，可能是认识到某一特定人口中存在大量阅读技能缺乏的人。类似地，项目之所以在实施中，是因为已经认识到了某种社会问题的存在：中学里的驾驶教育受到公众支持，因为有驾照的人发生车祸的比率很高。

所以，评估的重要方式就是，评估某个社会问题的性质、重要性和分布状态，评估针对这一问题实施干预的必要性，分析现有环境对干预的概念化和干预设计的意义。在评估领域，这些诊断行为通常被称为**需求评估**(Needs evaluation)，需求评估与其他领域中被称为社会流行病学和社会指标的研究是重叠的(McKillip，1987；Reviere et al.，1996；Soriano，1995；Witkin and Altschuld，

1995)。需求评估通常被作为设计和规划新项目或重组既有项目的第一步,用来提供需要什么服务和怎样最有效地向需要服务的人送达服务的信息。对已经建立的、稳定的项目来说,需求评估也很重要,能够检验服务是否满足了目标群体的实际需要,从而为项目改进提供指导。通过观察,需求评估可以了解潜在目标人群的需求,例如专栏 2—J 就展示了可以采用的几种方法之一。本书第 4 章将详细讨论需求评估的各个方面。

专栏 2—J　成年男女无家可归者对帮助的需求

在纽约市立庇护所,研究者调查了有代表性的 1 260 个无家可归的男女样本,要求每个成年人说出他们对需求的理解。这次调查有 20 项内容,每项内容说明一方面的帮助需求。大部分的回答提出了多元化的需求,平均 6.3 项。其中,提到最多的需求是安定的住所和稳定的收入,接下来是找到工作和提高工作技能。和女性比较,男性更多地希望摆脱酗酒、吸毒等问题的困扰,并希望学会合理使用金钱、领取退伍军人津贴、处理和警察的冲突以及与其他人友好相处,还有寻找住所。女性更经常提出的希望是健康和医疗问题、学会自我保护的技能。评估者指出,项目如果要对这些需求做出实际回应,就必须有能力送达或者通过机构提供大范围的服务。

资料来源:Daniel B. Herman, Elmer L. Struening, and Susan M. Barrow, "Self-Reported Needs for Help Among Homeless Men and Women," *Evaluation and Program Planning*, 1994, 17 (3): 249-256. Copyright © 1998, John Wiley & Sons, Inc.

项目理论评估

不管人们愿意不愿意,即使有了主要问题和干预的需求,也不见得会有项目来解决问题。项目的概念化和设计必须反应目标问题的基本假设,并提出有根据的、可行性的方法来解决问题。换句话说,每个社会项目都以某些计划或蓝图为基础,而计划或蓝图要依据那些对项目历史、目标和活动最理解的人的说法,提出有效的工作构想。尽管这样的计划很少作细节描写或者根本就不用写出来,但是,毋庸置疑,这样的计划是主要项目方共享的和默认的。因为这种项目方案的主要构成是项目实施和目标达成的一系列假设和预期,所以我们把它称为**项目理论**(Program theory)(下一章会有更充分的讨论)。如果理论不完善,那么,无论干预设计得多么精致或者实施得多好,都不会获得成功(Chen, 1990; Weiss, 1972)。

项目理论评估(Assessment of program theory)关注的是与项目的概念化和设计方法相关的问题。首先,项目理论评估包括以外化的和具体的书面或图表形式呈现项目理论。其次,可以用各种方法检验理论的合理性、可行性、道德性,甚至包括其他方面。项目理论评估是项目初期(在项目刚开始或调试阶段或规划阶段)基本的工作。而且,对于已经建立的项目而言,项目理论评估也是适用的,尤其是需要说明项目服务在何种程度上达到了他们尽量去满足的社会需求的时候。这种评估的主办方,通常是那些力图发展新项目的人,譬如基金机构或管理

者或希望确保项目概念化和设计适合于项目目标的那些人。例如专栏2—K,就描述了家庭维持项目概念的基本检验过程,说明基于这一概念框架的项目没有成功的希望。第5章对项目理论做了进一步的讨论,并讨论了项目理论评估的一些方法。

专栏2—K　家庭维持项目的设计缺陷

作为可评估性评价的一部分(见第5章),美国健康与人类服务部(U. S. Department of Health and Human Services)的评估者回顾了家庭维持项目(FPPs)的设计。FPPs是有时间限制的、集中在家庭基础上的、为那些处于危机状态下的家庭提供的服务,目的是使儿童处于被看护状态。评估者与联邦以及私营机构讨论了FPPs的定义,回顾了有用的文化背景,获得了政府和各地项目的描述,并且对4个项目点作了深入访谈。从这些信息中,他们提出了项目的"模式",假定项目是如何操作的,接着如何获得政策制定者、项目管理者和操作层面员工关于以下4个关键方面的意见:①项目目标,②对项目有影响的儿童福利体系的各个方面,③目标人群,以及④FPPs区别于其他建立在家庭基础上的服务特点。根据他们自己的分析和与专家委员会的讨论,评估者得出了结论:正如最近的分析结果所示,家庭维持项目没有达到政策制定者最初的目标(即把家庭作为看护场所)。项目设计的主要缺陷是难以识别处于"高危"状态的儿童;这就意味着,项目不能针对真正处于危机状态的目标家庭和儿童开展工作。

资料来源:Joseph S. Wholey, "Assessing the Feasibility and Likely Usefulness of Evaluation," in *Handbook of Practical Program Evaluation*, eds. J. S. Wholey, H. P. Hatry, and K. E. Newcomer (San Francisco: Jossey-Bass, 1994), pp. 29-31. Wholey's account, in turn, is based on Kaye and Bell (1993).

项目过程评估

即使知道了如何顺利地干预以便改善经过精确诊断的社会问题,项目还必须谨慎地实施才能合情合理地实际作用于目标问题。很多情况下,许多项目往往不能按照预先设计实施和执行。由于政治冲突,项目管理可能很粗糙或者不得不做出妥协。有时候,要么没有可用的项目人员,要么没有可用的设备;有时候,项目成员因为缺乏动力或经验,不能执行项目。经常出现的情况还有:项目设计有问题,进而为随意解释留下了很大的空间;或者没有很好地给项目工作人员转达项目的基本意图,使得项目活动一次次地拖延。还有,在项目所预想的参与者方面出现问题,要么参与者数量不够,要么不能准确识别参与者,或者参与者不合作。

所以,一个重要的和广泛使用的评估方式即**项目过程评估**就是评估项目实际实施的程度和效果。执行评估就是评估项目的过程、活动和项目的操作状况,因此,通常称之为**过程评估**。如果项目正在进行中,则称之为**项目督导**。过程评估致力于解决那些和项目如何运作相关的问题,包括服务和项目目标的一致程度、服务送达干预对象的好坏、项目管理的绩效、项目资源的使用情况以及对其他类似方面的问题(专栏2—L是一个范例)。

专栏 2—1 第一线的失败:实施福利改革

工作报酬(work pays)是加州的一个政府福利改革示范项目,目的是激励人们去工作,同时又阻止人们依赖于"家有小儿帮助"项目(AFDC)。项目管理者意识到,为了实现政策制定者的意图,当地福利办公室的工作人员本来应该将新政策告知项目对象,并且以积极的、个性化的方式,增进项目对象对劳动(工作)与福利(义务和选择)的了解,因而实施了执行评估。研究者就工作报酬项目调查了福利办公室的工作人员,并观察了大量的项目对象。这一信息揭示,在预设的新政策下,福利办公室工作人员和项目对象之间的交流方式非常少。对项目对象的调查显示,超过80%的工作人员并没有提供新政策的相关信息和对新政策进行解释。大多数工作人员只是按照原有的工作方式,搜集和核实项目对象接受项目是否合格的信息(资格审查)和向对象原文复述福利规定。但是,评估者也发现,工作人员自己掌握的关于工作报酬的项目信息量也非常少,也没有额外的时间和资源就项目的大量变动对项目对象进行培训。评估结果证明,在加州的街道层次,福利改革项目并没有得到很好的贯彻执行,也揭示了出现这种情况的原因。

资料来源:Marcia K. Meyers, Bonnie Glaser, and Karin MacDonald, "On the Front Lines of Welfare Delivery: Are Workers Implementing Policy Reforms?" *Journal of Policy Analysis and Management*, 1998, 17(1): 1-22.

项目过程评估是一种最经常用到的项目评估形式。过程评估常常用于对社会项目的评估,既可以用作独立的评估方法,也可以与影响评估一起用作比较复杂的综合评估的一部分(参见下面的讨论)。作为独立评估,过程评估能够提供高质量的信息。也就是说,过程评估可以根据既有的标准,评估项目按照预先设计实施的程度。当项目使用的是已有的有效模式之一,并且项目得到了很好的实施,那么,就有证据预期全面实施的结果。如果是新项目,过程评估就向管理者和其他项目方提供宝贵的反馈,即项目理论操作化的进展。从管理的角度来讲,过程评估提供的反馈可以使项目管理达到更高的水平(Wholey and Hatry, 1992),可以用 MIS 制度化的方式来搜集相关资料和主要指标,以便提供例行的、正在实施过程的反馈。

过程评估也有其他的一般性应用,对影响评估来说,过程评估是不可缺少的。影响评估提供的是项目产出信息,但是,如果没有关于导致项目产出的项目活动服务方面的信息,那么,对产出的说明就是不完整的和不明确的。当还没有发现项目影响时,很明显,过程评估也有诊断价值,通过过程评估来确定没有预期产出是因为执行失败(也就是说,并没有提供预定的服务,因此预期的效果也不会产生);或者是因为理论失败(也就是说,项目是按照预定方式执行的,但是没有产生预期的效果)。另一方面,如果项目产生了效果,过程评估则有助于确认项目产出是否是项目活动的效果,而不是虚假的关系,而且有助于识别与项目效果直接关联的项目服务,以便项目管理者了解他们要进一步努力的重点。本书第6章将对过程评估及其变种作更深入的探讨。

影响评估

影响评估(Impact evaluation),有时也称为产出评估,用来评价在一定社会环

境中的项目产生了哪些预先设想的对环境的改进。影响评估的基本问题涉及如下内容:是否获得了预期的产出,项目对社会环境的干预是否发生了作用,项目影响中是否包含意想不到的效果。

评价某个具体项目影响的最大困难,就是其预期产出往往是由与项目不相关的因素所造成的。影响评估的基本目标是对某项干预活动的净效果进行估计,也就是说,估计在没有其他过程和事件的影响下干预的纯粹效果。问题是,有些过程和事件同样作用于项目力图改善的环境。要实施影响评估,评估者就需要制订搜集资料的计划,运用所获得的资料证明观察到的变化是干预的结果,而且通过其他方式无法获得这样的结果。要做到这一点,需要对描述项目发生作用的产出变量作详细的说明,需要对这些变量进行精确测量,还要制订研究方案,方案不仅用上述的测量标准表达项目对象的状况,而且要估计出假如他们不接受干预的话,他们的状况将会怎样。影响评估的复杂性很大程度上是和对不接受干预者状况的正确评估相关的,就是所谓的反事实,因为它描述了与实际作用于项目对象的产出相反的一种情况(专栏 2—M 描述了这种情况)。

专栏 2—M　对减少垃圾没有影响

台湾是一个人口密度很高的小岛,受到垃圾问题的困扰。近年来垃圾成指数增长,垃圾处理点的数量明显不足。结果,1993 年在内府,台北的一个郊区,实施了一个示范性的垃圾减少项目(GRD),并且评估了项目对控制垃圾数量的结果。在台湾,垃圾每天都要搜集,GRD 计划打算通过暂停星期二的搜集工作破坏这种惯例。这样做的原理是,让居民每星期在自己家里存储一次垃圾,因为家里很少有这种存储职能,所以这样做会产生极大的不便,项目就是要通过这一点来增强居民对垃圾问题的意识。结果,可以预见到居民会努力减少每天制造的垃圾数量。过程评估证明项目是按照计划执行的。

通过项目实施前 4 个月和实施项目期间 4 个月的长期观察,掌握内府和南港(一个类似的毗邻郊区)的每日垃圾量记录,实施影响评估。分析结果表明,项目期间与项目实施前相比,内府的垃圾量并没有减少,对照社区也是如此。这一结果表明,居民只是将星期二的垃圾保留一天然后在星期三扔掉,而项目对每星期产生的垃圾总量不产生影响。对居民的调查显示,项目理论是不正确的——他们并没有像预期的那样对垃圾存放在家里表现出不方便或不高兴的态度。

资料来源:Huey-Tsyh Chen, Juju C. S. Wang, and Lung-Ho Lin, "Evaluating the Process and Outcome of a Garbage Reduction Program in Taiwan," *Evaluation Review*, 1997, 21(1): 27-42.

对评估者来说,要决定什么时候适合影响评估以及这个时候应该使用什么样的评估方案,还需要经过大量的考验。一方面,评估主办方经常认为他们需要的是影响评估,而且这是判定项目是否取得预期效果的唯一方法。另一方面,影响评估本身的特点要求必须拥有专家、时间和资源,但限于刻板的项目操作过程,经常很难适当地施展出来。如果对产出信息的需求足以证明影响评估的花费和所付出的努力是正当的,则还有一个悬而未决的问题,就是项目环境是否适合做这样的评估。举例来说,对某个既没有很好地被结构化,又无法详细地描述

的项目产出进行讨论,几乎就没有意义。因此,影响评估最适用于成熟的、稳定的项目,因为这样的项目有经过详细诠释的项目模型和对评估结果的明确利用。本书第 7 章到第 10 章将会论及影响评估以及设计、实施影响评估的各种方法。

效率评估

除非项目拥有经过证明的产出,否则很难给出将项目执行或维持下去的理由,因此需要进行影响评估。但是只有产出信息通常是不够的,项目影响还必须经过和成本对比再做出判断。这在当前的政治气候下尤其重要,因为扶持社会项目的资源缩减了,而且项目之间对经费的竞争越来越激烈。有些项目由于与产出相比成本过高而得不到支持(专栏 2—N 提供了一个范例)。

效率评估(efficiency analysis)分析项目成本和绩效之间的对比关系,解决这类评估问题的两类密切关联的方法是:**成本—收益分析**(Cost-benefit analysis)和**成本—绩效分析**(Cost-effectiveness analysis)。这类评估的典型问题包括:"相对于付出的成本而言,项目是否产生了足够的收益?"以及"项目创造的收益是否比其他致力于相同目标的干预或服务送达系统所消耗的单位成本要低?"

专栏 2—N　为精神病患者提供社区治疗的成本—收益分析

如果能得到支持性的服务,患有精神疾病的人通常就能够留在社区居住而不是住在政府医院中。但是这样的社区治疗所付出的成本是不是比留在医院中要多呢?俄亥俄的一个研究小组对比了为政府证明有严重精神残障的人提供家庭津贴与个案管理的社区项目的成本和地方精神病院的住院病人所花费的成本。在两年多的时间里,每个月都调查项目对象,了解他们关于精神健康服务、医疗和牙科服务、住房服务的消费,以及其他个体消费。有关这些服务成本的信息是从各自的服务提供者那里获得,并且与社区项目本身的直接成本联系起来。病人住院 90 天或更长时间所用的成本是从俄亥俄精神健康部门的预算资料中获得的,并且尽可能地按照为社区项目参与者提供的列表划分为不同的类别。然而,全面来说,所有服务的成本估计在社区项目服务中为每月 1 730 美元,而住院病人每个月大约是 6 250 美元。所以社区照顾的成本要比医院照顾少得多,而不是多得多。

资料来源:George C. Galster, Timothy F. Champney, and Yolonda Williams, "Costs of Caring for Persons With Long-Term Mental Illness in Alternative Residential Settings," *Evaluation and Program Planning*, 1994, 17(3): 239-348.

由于要假设项目相关活动的货币价值,有时候还要将项目收益转换为货币价值,所以效率评估很微妙,也颇具争议。但无论如何,效率评估对于确定项目资源的分配、识别同样经费条件下获得最大收益的项目模式、确定给某个项目以政治支持的程度,都是非常重要的。

和影响评估一样,效率评估最适用于成熟稳定的项目,稳定的项目具有高度组织化的项目模式。另外,正如上面提到的那样,效率分析也建立在过程评估和影响评估的基础上,所以项目影响的本质和范围应该先于或者同时与效率评估确定,这一点是很重要的。即使被要求实施效率评估的专家已经到位,很明显的是,只有当

需要而且使用这一信息的人存在时，才能采用效率评估。即使有对项目成本在多种背景中的高水平关注，这也可以被认为是一种普遍的事实。用于效率评估的程序虽然不如影响评估对资源和项目合作的要求那样严格，但是其程序高度技术化，而且需要高水平的专家。第 11 章将更详细地论述效率评估方法。

小 结

- 每项评估都必须经过准备来适应被评估项目的环境，以便于评估方案能够为指定的问题提供可信的、有效的解决方法，同时还有利用所掌握的资源来实施评估的充分可行性。
- 评估规划本身围绕三个主题建构：①评估要解决的问题；②解决这些问题所运用的方法和程序；③评估过程中评估者—项目方关系和互动的性质。
- 在一项评估规划所应该考虑的评估背景中，有三个基本方面：评估的目的；被评估项目的结构和条件；评估可资利用的资源。
- 影响评估规划的一个重要因素是评估意图达成的目标。评估的全部目标必然决定其重点、范围和组成。一般来说发起评估的目的，或为改进项目而向项目管理者和主办方提供反馈意见，或为向决策者就项目是否有效做出证明，或为某种社会干预方式的知识积累做出贡献。
- 影响评估规划易于改变的另一个重要因素是项目结构和环境的性质。评估方案必须反映出项目的进展阶段或程度，项目各方有关项目性质和任务一致或冲突的程度，项目基本原理和方案固有的价值及概念，以及项目组织和管理的方式。
- 评估规划也必须适应评估所能获得资源的限制。关键资源不仅包括资金，也包括允许实施评估的时间、相关的技术专家、项目和项目各方的合作、以及获得重要记录和项目资料的途径。从评估角度看，最值得做的和从可用资源角度来说最可行的，往往会有差别，评估者必须在两者之间找到平衡。
- 确定评估者和评估主办方、其他主要项目方之间的恰当关系，是经常被忽视的问题，同时也是评估规划的关键方面。有三种类型的项目评估各方关系，即①独立评估：经常被使用，评估者对设计和实施评估承担主要责任；②参与性或合作性评估：作为协作性的项目评估，评估者和部分项目方的代表以项目小组形式进行合作，项目各方之间进行更富有参与性或合作性的互动；③授权评估：评估的目的是帮助提高项目参与各方的能力，这样做可以增强他们的技术或提升政治影响。
- 在计划中确定的评估问题和解决这些问题的方法，一般分成一种或一种以上的可识别的类别，涉及①对服务的需求（需求评估）；②项目概念化和方案（项目理论评估）；③项目执行过程（项目过程，过程评估或者项目督导）；④项目产出（影响评估）或者⑤项目效率（效率评估）。在实践中，大部分评估方案都是按评估要解决问题的类型来选择评估方法，然后根据项目环境的特殊性进行针对性调整。

基本概念

项目过程评估(Assessment of program process):一种评估研究,用以回答项目的操作、实施以及服务送达等问题。亦即过程评估或执行评估。

项目理论评估(Assessment of program theory):一种评估研究,用以回答项目的概念化和设计等问题。

成本—收益分析(Cost-benefit analysis):确定项目经济效率的分析过程,表达成本与项目产出之间的对比关系,通常用货币形式表示。

成本—绩效分析(Cost-effectiveness analysis):确定项目绩效的分析过程,测量的是相对于项目成本而言,干预所获得的效果。

效率评估(Efficiency assessment):一种评估研究,用以回答项目成本问题,既要比较项目收益的货币价值,也要比较项目给社会条件带来改善的绩效。

授权评估(Empowerment evaluation):一种参与性或合作性的评估,评估者的角色就是为项目各方自己的评估提供直接帮助和咨询,使其更有效地做出支持或改进的决策、改变他们的项目、发挥他们的影响力。

评估问题(Evaluation questions):由评估者、主办方或其他项目方提出的一系列问题。这些问题将决定评估调查的主题,并且按照一定的方式表述出来,既能满足评估者的方法要求,又能为项目各方所用。

塑造性评估(Formative evaluation):提供项目改进方案的评估活动。

影响评估(Impact evaluation):一种评估研究,用以回答项目产出以及项目意图改变的社会状况等问题。亦即影响评估或结果评估。

独立评估(Independent evaluation):评估者全权负责制订评估方案、实施评估以及发布评估结果的评估活动。

需求评估(Needs evaluation):一种评估研究,用以回答项目运作所需的社会条件以及项目需求程度等问题。

参与性或合作性评估(Participatory or collaborative evaluation):作为集体性项目的评估,评估者和部分项目方的代表合作,制订评估方案、实施评估、发布和利用评估结果。

过程评估(Process evaluation):一种项目督导形式,用于判断项目是否按照预期把服务送达给项目对象。这也就是所说的"执行评估"。

项目督导(Program monitoring):一种核查各个方面项目绩效的系统过程,用来核查项目是否按照预期设想或者恰当的项目标准发挥了功效。项目督导涉及项目绩效的多个方面,一般包括项目过程、项目产出中的至少一个。

项目理论(Program theory):用于描述项目产生预期社会收益以及为完成项目目标所采取的策略和行动之间关系的一系列假设。项目理论可以分为影响理论(impact theory)和过程理论(process theory)。影响理论用来描述项目行动给社会环境带来的变化,而过程理论揭示项目的组织计划和服务利用计划。

总结性评估(Summative evaluation):对项目实施的某些关键部分提出综合性评价的评估活动,例如评估特定的目标和客观效果是否一致。

对象(Target):项目所指向的单位(个体、家庭、团体,等等),凡是项目所涉及的单位就构成对象群。

确定议题和设定问题 3

在先前的章节中，已经介绍了在制订评估标准时人们对所要考虑到的许多问题的总看法。尽管那些问题对于评估设计非常重要，然而评估规划的核心是要得到有关社会项目绩效问题的可靠答案。优秀的评估问题必须紧扣项目实质，使其能够完全揭示项目的本质，并且围绕着重要项目方所关心的问题展开。这些问题必须让评估者能用自己所熟悉的研究技能、方式来回答，同时还必须得到明确的阐述，进而能够明确地表达或直接地表达评估项目绩效所需达到的标准。

因此，一套合适的评估问题是整个评估过程的核心。对问题详尽仔细的阐述，能在很大程度上帮助评估设计，以使评估发现得到有效的使用。评估问题有不同的表达方式，对于项目各方和项目决策者而言，有些问题较之其他的更为有用。此外，评估问题的某些形式对评估者获得可靠的答案起着至关重要的作用；而且，一些评估问题的表达形式对项目绩效问题的考核比其他形式来得更加直接。

本章所介绍的是评估者获得有效评估问题的途径。为了达到这个目的，一个核心的过程就是识别使用评估结果的决策者，了解他们所需要的信息以及他们使用这些信息的方式。评估者个人对项目的分析也十分重要。实现这个目标的一个十分有效的独特方法，就是清晰地掌握项目理论，即对项目工作的方式和原因有详尽的说明。对项目理论的思考主要集中在那些对评估会产生影响的重要事件以及前提上。

在评估中,一个至关重要的方面就是对评估所涉及的问题进行识别和阐述。某些人或许会认为这个步骤十分简单,而且,这些问题只是一些在执行评估的过程中所形成的、习惯成自然的步骤。然而,正像在第 2 章里介绍的那样,在评估的最初,评估主办方对评估问题的具体描述对于最终形成可操作的评估问题特别重要。评估者也不能单凭自己个人的专业知识来决定要关注的问题,这种行为只可能导致危机。在这种情况下,评估将无法解决项目各方所关心的问题,进而使评估变得毫无意义,而且很有可能被攻击为文不对题。

为了确保评估能够直接切入相关决策者和项目各方最关心的问题,在做最初的评估时,最好与决策者和项目各方进行交流、商议,之后,再将评估问题列举出来。然而同样重要的是,在为评估下定义的过程中,要与重要的项目方接触以增进彼此间的共识,当得到评估结果时,这些项目方便能理解、欣赏和充分利用评估发现。

尽管项目各方的参与非常重要,但是,评估者不能单纯地依赖决策者和项目各方去分析评估将要涉及的问题。有时,评估主办方对评估本身十分有经验,并具备做必要背景工作和阐述详尽可行问题以及分析与评估相关问题的能力。然而,更常见的情况是,评估主办方和项目各方对评估的专业知识并不了解,如果是这样的话,他们就无法为评估准备背景工作。这将意味着,在有用的、通俗易懂的和详尽的评估结果出来之前,评估者无法得到一份完整的涉及评估问题的清单。这样,就无法为评估者提供既有的重要问题,也就不能使他们直接将这些材料转换成研究设计。

因此,在设定评估问题上,评估者也扮演着重要的角色。在面对项目的实践性和政策性问题时,项目各方是专家,但评估者最清楚如何对项目进行分析以及找到评估的关键问题。因此,评估者必须随时准备提出那些有可能被忽视的问题,确定项目执行过程中可能出现问题的各个方面以及需要调查的结果,此外,评估者还得和项目各方合作,将后者所关心的问题转化成可评估问题,并将问题组织成能被人们回答的形式。

在通常情况下,评估者较为明智的做法是将细节性的并能指导评估设计的问题用书面形式表达出来。这样,在设计可行性评估和选择有用的研究步骤时,便有据可循。更为重要的原因或许在于,这种书面陈述能帮助评估者与评估主办方和重要的项目方商议问题,来确保评估能包括后者所关心的问题。这个过程同时还能确保在评估执行过程中,人们不会产生误解。

本章节的以下部分将从两个重要方面入手,内容涉及如何获得能指导评估执行的问题:①以何种方式来形成与表述评估问题,从而确保评估者能够利用研究程序及方法对这些问题进行处理。②如何确定评估将要关注的具体问题。

获得好的评估问题的条件是什么

评估问题的形式应该由其必须完成的功能来决定。评估问题的主要角色是

将评估重点放在对主要决策者和项目各方而言十分重要的项目绩效上，同时能够推进对资料搜集程序的设计，从而为项目正常发挥绩效提供有意义的信息。因此，优良的评估问题必须清楚地确定测量项目绩效好坏的尺度，这样，评估出来的绩效才可靠。由此看来，这样的评估需要对绩效本质和评估标准有精确的描述（参见专栏3—A）。因此，好的评估问题必须详细地说明用来测量项目绩效的尺度，从而使项目绩效的测量有据可循。把这些不同的方面结合起来，才能保证进一步讨论的正确方向。

专栏3—A 评估的意义

在实践领域，调查的方式多种多样，比如在法律、医学和科学领域中都有此类做法。对于每一种调查而言，都有一种通用的形式帮助人们找到实践的基本逻辑，实现对实践的指导……评估是调查的一种，同时，也有一种基本逻辑或通用模式，让人们实现调查实践（斯克莱文（Michael Scriven）已对此有所研究，并取得了一定成就）……评估的基本逻辑是这样的：

1. 确定价值尺度：在什么样的维度上进行评估才能对被评估对象进行准确评估？
2. 制订标准：被评估对象在工作时所要达到的标准是什么？
3. 测量绩效并与标准相比较：被评估对象的表现如何？
4. 将资料综合并进行判断：被评估对象的优点或价值何在？

……评估某件事物意味着将其优点或价值与某些尺度和标准相比较。斯克莱文所阐述的基本逻辑清晰地反映了我们进行评估的过程。

资料来源：Deborah M. Fournier, *Establishing Evaluative Conclusions: A Distinction Between General and Working Logic*, New Directions for Evaluation, no. 68 (San Francisco: Jossey-Bass, 1995), p. 16.

项目绩效的维度

好的评估问题首先必须是合理并且适宜的。这就意味着，评估问题能满足项目各方对不同维度的绩效要求，并能同时提供实现项目操作的领域。但是不能把绩效问题想当然地认为是很容易的事情。例如，在低收入地区实施住房改革项目时，如果要求该项目使那些日益风行的毒品交易有所缓解的话，显然是不切实际的。同样，在进行办公室文件柜采购时，要求该项目能帮助此交易顺利地进行也是不合时宜的。进一步而言，评估问题必须是可回答的，这就是说，它们所包含的绩效维度必须足够具体、实际、可操作、可测量，这样才能得到有意义的绩效信息。如果要评估者判断一个成人读写能力项目是否提高了某一区域在世界经济中的竞争力，或让他去判断一个毒品预防项目中的顾问是否充分地照顾到了他与客户之间的关系，就显然有很大难度。

评估问题必须是合理并且适宜的

项目的鼓吹者常常会提出宏伟的目标（例如，提高和改善孩子们的生活质量），并期盼发生一些不切实际的结果，或者相信项目能实现一些与其实际能力

不符的成就。好的评估问题包含的绩效维度对项目本身而言应该是切合实际的。这意味着评估者必须经常与项目各方合作,这样便能准确掌握项目规模,从而将精力集中在评估问题上。例如,一位健康项目的经理在最初会这样问,"对于艾滋病威胁人类这个问题,我们的教育和服务活动是否已经成功地教育了公众?"而在实践过程中,这些服务或许仅仅只包含为数不多的健康教育活动。通过这些微不足道的活动,要公众最大程度地了解有关艾滋病的信息似乎是不可能的,更不用提用这些信息来帮助降低艾滋病对公众的威胁程度了。如果关于此类服务的问题对评估十分重要的话,那么一种更好的询问方式应该是,"在有关艾滋病问题的宣传过程中,我们的教育和服务是否让听众对艾滋病有了一些认识?"以及"听众是否主要是那些能在很大程度上影响其他人观点的社区领袖?"

还有两种方法能够帮助评估者来与项目各方进行合作,以判断候选评估问题与现实相符合的程度。其一是,在相关项目活动场景中去检测问题。比如,在上面的例子中,很显然,那些没有达到要求的教育和服务活动并没有完成所规定的任务,即"向公众介绍有关艾滋病对人类威胁的问题",所以当人们在判断任务是否已完成的时候,便毫无头绪了。评估者和项目各方应该确定和核查项目的构成、活动以及人员安排,这些都与项目绩效和评估问题阐述息息相关,同时,他们应该以一种合理的方式来形成有关以上特征的评估问题。

另一种评价候选评估问题的方法是,与应用社会科学和社会服务既有的经验和发现相比较,将评估与其联系起来,并进行分析。例如,一个有关青少年罪犯项目的评估主办方最初要获得的信息是,项目是否唤醒了罪犯的自尊,因为项目的信念是,对于这些青少年罪犯而言,自尊心的提高能帮助他们改良自己的行为。然而,应用社会科学研究表明,在通常情况下,青少年罪犯在自尊心方面并没有问题,而且事实证明,自尊心的提高并没有有效地减少青少年的犯罪情况。这些信息表明,评估者和评估主办方不得不承认,有关自尊心项目影响问题根本就是不恰当的。

阐述合适而又现实的评估问题,首先要详尽而完整地描述项目。在项目评估早期,评估者应该对项目有一个彻底的了解——项目构成如何、要举行怎样的活动、不同人员的角色和任务如何、参与人员的种类以及对项目重要功能的假设。与评估者合作的项目各方(尤其是项目经理和项目工作人员)当然也会对项目有所了解。在对项目活动和假设进行完整的思考之后,总结出来的评估问题自然就会与现实更符合了。

评估问题必须是可以回答的

很显然,由评估规划所构建的评估问题应该是可以回答的。不能回答的问题或许能引起哲学家的极大兴趣,但是对于那些需要获得评估结果的评估者和决策者而言,显然毫无意义。在尚未对整个事态完全掌握之前,要组织一份可以回答的评估问题的难易程度如何,或许是一个难以确定的问题。这种情况是有

可能发生的，因为有时候，在问题中使用的术语或许十分大众化，然而当需要有明确的概念时，便显得含糊不清了（譬如“这个项目是否提高了家庭的价值？”）。或者有些听起来十分合情合理的问题，在操作上清晰度仍然不够。这样，根据这些问题并不能获得什么（譬如，“根据情形分析，经理对他们客户所处的社会环境敏感吗？”）。同样，有些问题缺乏对答案相关尺度的充分提示（譬如，“这个项目成功吗？”）。最后，有些问题或许可以回答，但前提是，如果要回答这些问题，人们需要拥有更多的专业知识、资料，或者是资料来源，而在目前所拥有的信息远远不足以回答（譬如，“这个项目提供给高危女性的产前服务，是否能够帮助她们的孩子拥有更多的机会完成大学学业？”）。

为了让一个评估问题能够得到回答，必须事先确定一些依据或搜集一些现实的信息，使人们能在可靠的基础上回答问题。总之，这意味着要总结出一些用清晰并且没有矛盾的定义表述出来的、关于绩效的、可测量的问题。另外，相关的测量标准也必须同样明确而具体。例如，我们设想在早期教育（head start）项目中，有这样一个有关教育经费补贴的评估问题，“我们的项目是否在最大的程度上满足了孩子们的需求？”为了增强该问题的可回答性，评估者应该能够做如下拓展：

A. 确定该事件中特定的孩子群（例如，人口普查中年龄在 5 岁及以下的孩子，家庭年平均收入低于联邦贫困水平 150% 的孩子）；
B. 确定能反映最大需求的具体测量特征和价值（例如，低于联邦贫困水平的年收入额，单亲家庭中被剥夺了接受低于高中教育的孩子数）；
C. 举出有可能导致的评估发现（例如，在项目内，目前有 60% 需要接受教育的孩子还未接受教育，在**覆盖区域**（Catchment area）——项目所涉及的地理区域中需要接受教育的孩子中有 75% 没有被纳入项目）；
D. 规定评估标准（例如，为了使项目产出令人满意，至少有 90% 的孩子需要接受教育，而项目区的孩子至少要有 50% 被纳入项目）；
E. 让评估主办方和其他项目方（这些人将参与评估的整个过程）认可：满足这些标准的评估发现确实能够回答所提出的问题。

如果这些要求可以达到，同时，有条件可以随时进行搜集、分析和呈现可用资料的话，那么评估问题理所当然就具有可回答性了。

项目绩效的标准

从合理的具有可答性的问题开始进行研究，是社会科学的一个惯例（尽管经常以假设的方式呈现问题）。评估问题的特殊性在于其与绩效息息相关，至少与一些可判断的绩效标准紧密结合在一起。前面虽已经提到过确定相关标准是使评估问题具有可答性的一个重要步骤，但是考虑到这个议题在评估问题中的重要性与特殊性，我们将单独对其进行讨论。

当项目经理或是评估主办方询问这样的问题（如“我们所选定的服务群是否准确？”或者“我们的服务是否为项目对象带来了收益？”）时，这说明他们不仅询

问了有关项目服务于特定人群的情况和提供服务所带来的收益,还说明他们也关心这些绩效是否能达到一定的水平和标准。毫无疑问,至少有一些"适当的项目对象"接受了服务,或者有一些项目对象从服务中获益。但这些是否就已经足够了呢? 一定的水平或标准必须用数字和数量的形式表达出来,从而实现在这些维度上对绩效的评估。

评估的这种特征意味着:好的评估问题,如果可能,在表达可用的**绩效标准**(Performance criterion)的同时,还能提供该议题上有关绩效的讨论尺度。这样的话,评估问题应该是这样的:"在接受项目服务的对象中,是否至少有75%的人适合该服务?"(对适合与否得有一个清楚的定义)或者"接受了职业咨询服务的人,是否大部分在接受了3个月的职业培训之后,在30天之内就找到了工作?"同时,在这些问题中表现出的绩效标准尽管有些并不是很直接,但却与项目所要服务的社会需求相联系。要达到那样的标准是有原因的,根本原因是能帮助项目充分地将其绩效发挥出来,以达到要求的水平,从而达到改善既定社会状况的目的。

对于评估者而言,一个值得考虑的因素是,不同项目绩效维度可用的绩效标准有多种(参见专栏3—B),而事实上,在搜集资料和汇报结果之前,并不总能获得明确而又能被公认的绩效标准。虽然如此,在某种程度上,如果初期评估问题的阐述就包括清楚的绩效标准,如果这些标准得到主要项目方的认同,评估规划的实现将会变得相对容易,同时也减少了在介绍评估结果时收到相反意见的潜在可能。

专栏3—B 可能与项目绩效有关的多项指标

在一项评估中,项目绩效的标准包括:

- 目标人群的需求或需要
- 陈述出来的目标和目的
- 职业标准
- 惯例性实践;其他项目的规范
- 法律要求
- 伦理或道德价值;社会正义与公平
- 过去的绩效;历史资料
- 由项目经理设定的目标
- 专家意见
- 目标人群干预前的基准
- 在缺乏项目帮助的情况下会出现的情况
- 成本或相对成本

所以,我们强调绩效标准问题的不可回避性。而一个仅仅只描述项目绩效的评估不是一个真正的评估(定义参见专栏3—A),至多能帮助人们设定标准和

判断在信息使用过程中的绩效表现。

在考虑到了这些问题之后，我们将把注意力集中到不同种类的绩效标准上，这些绩效标准与阐述有意义的评估问题相关。或许最为常见的标准是那些基于项目目标和目的之上的标准。因此，某些可预期的成绩就成为项目管理人员和主办方所确定的项目目标。在通常情况下，如果考虑项目绩效的本质和水平，那么，这些目的和目标的陈述就不是非常细致（本章将会讲到，评估者应当区分项目的总体目的和具体的可测量的目标）。例如，为受虐待的妇女提供庇护处所的项目，其中的一个目的就是“使她们能够有能力主宰自己的生活”。尽管这反映了某些值得表扬的价值观念，但是这种陈述并未给出测量这种自主能力的具体指标，也未指明何种水平的增权才算是实现了这一目标。在将粗略陈述转译成通俗易懂的术语体系的过程中，与项目各方适当的商讨是十分必要的，只有能被大家接受的术语才能将需要达成的项目产出更加具体地描述出来，才能确定与这些产出相当的、可观测的指标，才能区分成功完成预定目标的程度和层次。

然而，有些项目的目标十分具体。当行政目标被作为常规项目功能的目标时，常常会发生这种情况。在设定目标层次时，人们往往会考虑过去的或从相似项目中得到的经验，还会考虑某些合理或者合意的判断，抑或仅仅是“最佳猜测”基础上的判断。行政目标的例子有，在30天以内，使被推荐的人90%进入项目，完成对75%客户的服务，在客户填写的问卷上有85%表示“好”或“很优秀”，（在这种管理方式下）得为每个人提供至少3种适当的服务，等等。很显然，这些标准的层次的确定存在一定的任意性，但是如果能用行政性方式规定下来，或者在项目各方共识的基础上具有合理性，从而被决定下来，那么在阐述评估问题和对其后续发现进行解释时，这些标准就很有意义。然而，如果项目没有能力做出详尽陈述，而评估者却坚持要作出绩效目标的精确陈述，就不是明智之举。将目标定在一个较高而又任意的基础之上，会导致评估标准在不同情况下被任意地修改。

在一些情况下，有一些既有的专业性标准可供人们选择并作为绩效评价的标准。在医疗和保健项目中，这种情况较为突出，其中已有许多实践指南和护理水平标准模式，在此类项目确定绩效标准时，这些标准便能起到指导作用。然而更多的情况是，在没有既定标准的情况下，或者是没有任意行政目标去调用时，这些指标便起作用了。典型的情况是，绩效标准本身就总结得十分清楚，但是依据该尺度测量的绩效标准还是具有不确定性。例如，项目各方或许会同意项目应该有较低的排他率、高的完整服务比例、对客户的满足程度有较高的水平，等等，但是仅仅有这些模糊的意见是不够的，因为按照单个尺度来看，什么样的水平算是“高”或“低”都无从确定。有时，评估者能够利用先前的经验，或者是在评估和项目文献中找到一些信息，从而为确定绩效标准水平提供合理的基础。另一种方法是，从项目各方那儿获取信息，按照对某种意见赞成的多寡来确定标准，或者确定标准水平的一个区间，而这个区间能识别出比方说高、中或低等的不同绩效。

在评估问题中,当绩效标准涉及项目产出或项目影响问题时,要确定标准就尤其困难。产出会与预测的出入有多大,对此问题项目各方和评估者毫无概念(例如,对于麻醉剂使用的尺度标准)。在缺乏标准的情况下,常常要基于统计学的标准进行判断。例如,项目被断定为有效的,是因为可测量的效果具有统计显著性。这是一种低"标准"的做法,在后面讨论影响评估时会对其具体原因做更为完整的考察。统计标准与产出维度变化的实际意义之间并没有内在联系,并且可能产生出误导。一个统计上具有显著效果的、使再犯率下降2%的青少年犯罪项目,也许实质上并没有多大意义和影响,不值得再继续实施。因此,评估者应该尽可能地使用上文所提到的技术,尝试确定和细化一个可操作的、合适的"成功"标准,来判断项目所获得效果的性质和程度。

典型的评估问题

从以上的讨论中不难看出,较好的评估问题在评估环境中往往十分具体、详尽。具体问题总是与具体的社会项目联系在一起,有些项目问题的变化不大,有的可能很大。正如第2章提到的,评估问题主要是处理五种综合性项目问题中的一种。每一类中比较一般的问题,按照概要的方式在下面罗列了出来。

有关项目服务需求的问题(需求评估):

- 问题的本质与范围是什么?
- 需求人群的特征是什么?
- 人群的需求是什么?
- 需要什么样的服务?
- 所需服务的规模多大,在什么时候需要?
- 为了将服务提供给人群,应该安排怎样的送达渠道?

有关项目概念化或设计的问题(项目理论评估):

- 应该为什么样的客户提供服务?
- 提供什么样的服务?
- 对服务而言,最好的送达渠道是什么?
- 项目怎样才能确定、重新招募和保证既有对象的数量?
- 应该如何组织项目?
- 对于项目而言,什么样的资源是必需而又合适的?

有关项目操作和服务送达的问题(项目过程评估):

- 达到了行政性和服务性目标吗?
- 既定人群得到了既定服务吗?
- 是否存在需要此类服务,但服务还未涉及的人员?
- 在服务过程中,是否针对足够数量的客户完成了服务项目?
- 客户对服务满意吗?

- 行政的、组织的以及个体的功能是否得到了充分发挥?

有关项目产出的问题(影响评估):

- 需要达到的目标和目的是否已经达到?
- 服务对参与者是否有有利的效果?
- 服务对参与者是否有负面的效果?
- 服务对某些参与者的影响是否比其他人的要大?
- 服务企图改善的问题或是情况是否有所改善?

有关项目经费和效率的问题(效率评估):

- 资源是否被充分利用?
- 与收益最大化比较,成本是否合理?
- 是否还有其他的方法能帮助降低成本并获得同样的结果?

评估者必须认识到,上述各类问题之间有着非常重要的关系。每一个问题的设定,都从对其前面问题的回答中获益良多。例如,关于项目的概念化和设计的问题,就在很大程度上取决于该项目想要达到的需求的本质。不同的需求可以通过不同类型的项目来获得最好的满足。如果一个项目关注的是经济资源匮乏的人群,合适的项目概念和评估问题将与减少酒后驾车的项目有所不同。而且,评价项目设计的最适宜的标准,关系到这个项目设计与需求以及需求群体的环境有多么切合。

同样,设定关于项目执行和服务提供的核心问题,关系到这个项目设计在多大程度上能具体实施。相应地,评估项目执行和服务提供的主要依据,是对项目设计初衷的忠实程度。因此,关键的评估问题,关系到项目的意图有多少在实际中得到了落实。这就意味着,评估项目实施的标准至少部分地由项目本应有怎样的实际作用这一初衷所确定。换言之,由其基本的概念化和设计所确定。如果我们知道一个项目的本意在于通过流动厨房(soup kitchen)为无家可归者提供食物,那么我们就会认识到——如果实际上并没有给无家可归者发放食物,那么项目一定是出了问题。

只有当项目设计得到了很好的实施,项目产出的问题才谈得上有意义。如果项目并没有真正提供某种服务,或者所提供的并不是本来意图提供的服务,产生预期的产出就是无稽之谈。当项目的实施并没有真正产生预期的改善,从而致使项目缺乏效果时,项目评估者称之为**执行失败**(Implementation failure)。如果由于流动厨房很少开放,使得没有多少无家可归者能真正在那里就餐,导致他们的营养状况并没有改善,这种情况就是执行失败。

如果体现在相应服务中的项目概念本身具有缺陷,即使一个项目被很好地执行,却也仍然有可能达不到预期的结果。如果项目的概念化和设计无论得到怎样良好的执行,都不能产生预期的产出,那么评估者就将这种情况称之为**理论失败**(Theory failure)。因此,在上述的评估问题序列中,那些关于项目结果的问题,只是在项目运作和服务能很好地执行时才有意义。同样,只有在确定了预期

行动和服务的项目设计表现出对社会问题以及项目旨在改变的需求的恰当回应时,好的执行才具有意义。如果流动厨房远离无家可归者聚集地,那么无论项目被执行得多好,都没有什么实际效用。在这个例子中,项目基本设计的一个关键方面(地域)并没有与目标人群的需求很好地关联。

在上面的列表中,关于项目经费和效率的最后一组评估问题,也与它们前面的各组问题有着显著的关联。除非项目已经产生了一点成果,否则,考虑产生这些成果所需成本或者获得这些成果的更有效方式的问题没有多少意义。举个例子,即便我们可以估算一下为无家可归者提供食物的流动厨房花费;但是,如果根本没有无家可归者需要提供食物,这一举动就毫无裨益,那么任何花费都可以说是过多了。较之于因不恰当执行或者项目设计缺陷所导致的无效活动,开支问题只是一个小问题。

评估的等级

上述各类不同评估问题之间的关联,设定了一个评估议题的等级结构,比之将各类问题简单组织起来,这一结构的意义更大。在第2章总体上介绍各类评估研究时已经说到,评估问题的类型以及回答它们的方法各不相同,每一种都构成一个自有其理的评估方式。这些评估问题和方法的分组,就像是构成整个评估研究的一块块积木,在每一个实际的评估研究中,我们都可以看到它们单独或者联合地发挥着作用。就像在专栏3—C中我们可以看到的那样,我们可以认为这些评估问题的积木搭建起了一个等级结构,每一个评估问题都以处于其下方的问题为依托。

专栏3—C 评估的等级结构

项目成本和效率评估(Assessment of Program Cost and Efficiency)

项目产出/影响评估(Assessment of Program Outcome/Impact)

项目过程和执行评估(Assessment of Program Process and Implementation)

项目设计和理论评估(Assessment of Program Design and Theory)

项目需求评估(Assessment of Need for the Program)

在评估的等级结构中,对项目需求的评估处在最基础的层次。对于社会问题的本质以及干预问题做出评估,产生出的诊断信息可以支撑有效的项目设计,也就是关于如何处理项目旨在改善的社会环境的项目理论。一旦有了一个可靠的项目理论,评估的下一步就是要考察项目执行情况如何,这正是过程与执行评估的任务所在。

如果我们已经恰当地理解了社会需求,应对问题的项目理论也是合理的,相应的项目活动和服务也得到了很好的执行,这时评估项目产出才变得有意义。因此,承担一项对产出进行评估的影响评估,有必要预设对在评估等级结构中低于这个问题的各个议题的评估都有可接受的结果。如果在影响评估完成时并没

有对前面的问题进行评估，那么，只有在关于前面问题的设定确实可靠的情况下，影响评估的结果才能得以解释。

对项目成本和效率的评估处在评估等级结构的顶端。对这些问题的评估需要对等级结构中的所有支撑问题都有了解。因为，只有在项目产出的性质、执行、理论和所关注的社会问题的信息都齐备的情况下，回答项目成本与效率问题才有意义。

因此，在确定项目评估问题时应该重点考虑的是，关于这个项目我们已经知道了什么或者可以确定什么。例如，对某个评估而言，如果对项目概念和执行状况都不明白，就将评估关注点放到对结果的评估上是没有什么意义的。

当提出评估计划的各项问题后，评估者最好从评估等级结构的最下层着手，并且思考：关于最基本的问题知道了一些什么，还需要知道什么。当确认可以完全接受各项前提后，也确定了需要回答的问题之后，再进入到等级结构的另一层级才是合宜的。依据对更基础问题的了解，评估者就可以确定当前层级的问题是否有意义。

通过牢记评估等级结构中各层级相应评估问题之间的逻辑相依性，评估者能够将评估的关注点集中在最恰当的项目问题之上。同时，也可以避免过早地关注更高层级评估问题类的错误。

确定评估应当回答的具体问题

项目的需求、设计、执行、产出和成本的各个评估问题并不是互斥的，一个评估所针对的项目问题并不只有一个类别。为了提出恰当的评估方案，对于项目可能涉及的许多问题，应限制在与项目环境关联最为紧密的具体问题上（关于一个实际项目的具体评估问题的例子，可以参考专栏 3—D）。

我们在第 2 章已经强调过，在评估方案各问题的形成过程中，要特别考虑评估主办方和其他主要项目方的关注点。在下面的讨论中，我们首先考察的是评估设计前和设计中，从评估主办方和主要相关项目方获取的问题。

然而，要评估者只将注意力集中在评估主办方和项目各方的身上来决定评估关心的重点，是不太恰当的。正因为他们对项目太熟悉了，项目各方很有可能忽视一些与项目有关的、重要的、但又很常见的方面。评估者所拥有的经验和知识或许会让项目获得新的切入点，他们的参与对确定相关评估问题也很重要。因此，从总体上看，评估者需要对项目做相对独立的分析，而这或许会对项目研究至关重要。因此下面将涉及的议题是：在评估设计过程中，评估者如何才能通过自己的方式分析项目，从而揭示潜在的重要评估问题。前文介绍的评估等级结构是实现这个目的的十分有效的工具。另一个特别有用的工具则是项目理论。通过对项目赖以成功的、对重要设想和期望进行的描述，项目理论能够将评估中应该关注的关键问题凸现出来。

呈现评估主办方和主要项目方的关注

在做计划和制订评估方案时，评估者常常发现自己会遇到一些持不同意见的项目方的反对，有时这些人对评估或项目的意见会与评估者的完全冲突，这是因为他们的利益有可能受到评估结果的影响（见专栏 3—E）。第 2 章已经讲过，评估者通常要与政策制定者、决策者、项目和评估主办方、目标群体、项目管理者和职员、项目竞争者、前后相关的项目方以及其他评估和研究群体打交道。在评估规划阶段，评估者往往试图确定各项目方提出的问题，根据问题的优先级别排序，并尽可能地将所有相关的关注点纳入到评估规划中来。

专栏 3—D　邻里课余活动项目的评估问题

在邻近的一个经济较为落后的地区，课余活动项目借助当地的一所小学为该区内一些无人照顾的孩子(latchkey children)在 3:00 到 6:00 之间提供免费的课后义务教育。这个项目的目标在于，为这些孩子们提供安全的、具有监督意义的学习环境，从而通过这种补课活动来提高他们的学习质量。以下是有关这个项目应该涉及的评估问题，仅以此作为例子，以供参考：

开展项目是否有必要？

问题：距该校方圆 1.5 英里的区域之内，有多少这样无人照顾的孩子？无人照顾的孩子的含义是，到了学习年龄但是在学习期间至少一个星期中就有一次没有成年人监护的孩子。

标准：在确定的区域内，应该至少有 100 名这样的孩子。而项目计划招收的孩子只有 60 名，这样可以充分保证每天进入项目的孩子能够达到足够的数量，因为种种原因，一些符合条件的孩子也许不一定进入项目。

问题：项目规定的、被包括进来的此类孩子的比例如何？

标准：进入项目的孩子中，至少有 75% 的应该完全符合项目条件。这是行政性目标，即在认识到其他孩子也有兴趣的情况下，进入项目的孩子主要应该是“无人照顾的孩子”。

项目设计的情况如何？

问题：计划中的教育活动对于这些孩子而言是否是最好的，是否能最有效地提高他们的学习效率？

标准：在对此类教育进行研究时，应该有指标说明这样的项目会有什么样的产出。同时，这些活动也应该得到相关年级有教育经验老师的认可。

问题：项目工作人员的数量是否足够？

标准：项目职员与学生的比率应该超过国家为正式儿童教育所规定的比率。

项目是否得到了有效执行？

问题：儿童参与的比例为多少？

标准：儿童的参与必须按照时间表的安排来活动，除非有家长的申请才能缺席。

问题：项目是否为学校的作业以及相关任务提供了正常的支持？

标准：孩子们必须在每天下午有平均 45 分钟的时间在有监督的情况下完成作业，还得阅读，这是每个参加的孩子必须完成的。

项目是否有既定的目的?

问题:参加的孩子对学校的态度是否有所改观?

标准:在整个学年里,必须有80%的学生对学校的态度有所改观。有资料显示,类似的孩子对学校的态度每况愈下;该项目的目的就是扭转这种局势,即使态度的改变并不大。

问题:孩子们进行了有规律的学习之后,他们的学业是否有所进步?

标准:孩子们在学业上所取得的平均成绩必须比没有参加该项目的孩子优秀。

项目的成本—效益比如何?

问题:在不计正常学年教育开销的情况下,每一个孩子参加此类项目的开销是多少?

标准:此开销应该和本州其他学区开展同样项目的相似或更少。

问题:除项目主管之外,如果项目工作人员是由社区志愿者而不是付酬的辅助专职人员组成的话,那么该项目是否能花销得更少而达到同样的效果?

标准:志愿者基础上的项目年开销,包括成员的组建、培训和志愿者的支持服务项目,将至少比目前开展的干预少花费20%的资金,而绩效并不降低。

出发点当然是评估主办方。这些评估的组织者兼资助人有设定问题的优先权。有时,评估主办方已经规定了评估问题,而且将解决问题的方法也研究出来了,他们仅仅需要评估者确定操作的具体细节。在这些情况下,评估者需要考察是否有项目各方的关注点被排除在评估主办方设定的问题之外,以及这些关注点是否足够特殊并且重要,以至于忽略它们将会影响到评估的质量。如果是这样的话,评估者必须确定是否要在特定的限制之下来执行评估,并在报告评估结论的同时表明这些局限以及偏差;还是试图对评估计划的安排展开协商,使得评估的范围扩大以容纳更多的关注点。

专栏3—E 关于无家可归人员的问题:在多机构项目中项目各方的不同评估意见

发起这个合作项目的目的在于,在城市、区域和大的组织中,以及在20个非营利公共机构中,改善对无家可归人员的健康与社会性服务。通过项目发展的服务包括指定性服务、自动服务中心、健康服务中心、医疗和护士服务中心以及项目管理。为了确保项目各方的参与,评估指导委员会将组织不同类型的代表参与到项目中来,同时,另外还会有两个项目方对项目负责。

尽管所有的项目方共同承担着一个责任,这也无可厚非(为无家可归人谋福利)。但对于评估而言,各方还是会有不同的意见。这一点在下文就有体现:

> 最不平衡的地方表现在不同部门有不同的组织文化,这样便导致他们在评估过程中表现出很大的不同。有些参与该项目的服务性部门持有相当前卫的评估意见,因为他们对这种服务了如指掌。他们十分了解整个评估过程,包括行政性程序和对义务的衡量。然而,在非营利部门,有些新入行者则以该领域改革者的身份希望评估能够重新恢复部门的活力并对改革提出了许多建议;其他非营利部门是宗教性或是慈善组织的分支,他们从事管理无家可归者工作已经有很长时间;对这些组织而言,评估(以及具有逻辑性的、以计划为基础的项目本身)则完

全超出了他们的工作经验。他们将评估者当做是一群好管闲事的人,所以直到现在,他们还是在非常不情愿的情况下尽量与评估者相处。他们最关心的是客户。除了公共部门外,他们会认为评估是浪费时间、金钱和能量的行为,类似这样的项目大多数以这样的结果而结束。他们被要求加入到整个工作的程序当中来,但是他们对此行却一无所知。

资料来源:Céline Mercier,"Participation in Stakeholder-Based Evaluation: A Case Study," *Evaluation and Program Planning*, 1997, 20(4): 467-475.

然而,评估主办方的最初意见往往并不具有很大的强制性,他们可以接受其他项目方的意见并进行协商。在这种情况下,评估者可在条件允许下尽可能地向所有的项目方咨询,将合情理的意见做优先的考虑,尽量做出一份能反映所有相关方面意见的计划。

考虑到项目各方及其意见的多样性,尽管评估者也投入了大量精力,在评估者和至少项目一方或多方之间还是会存在着产生对评估应关注的问题的不同理解。因此,在计划实行的最初,在评估者和项目各方之间进行充分而又坦诚的交流尤其重要。在搜集了项目各方对项目和评估的重要意见之后,与各方的交流会促使将评估重点放到现实中来,通过对项目的共同理解,确定评估所需完成的任务并阐明原因。最基本的是,评估者应该尽力确保最重要的项目方知道和接受评估的过程、评估将产生的信息类型、结果的意义以及可能留下哪些没有弄清楚的或没有得到回答的问题。

从项目各方那里了解信息

从定义上理解,主要项目方对评估和项目怀有浓厚的兴趣。这样,就需要识别这些人士并让他们说出评估应该涉及的方面和问题。评估主办方、项目执行者(或许本人也是评估主办方)以及项目的受益人,实际上常常是主要项目方。对其他重要项目方的识别通常可以通过分析项目所涉及的关系网来确定。最有启迪作用的关系包括项目的资金运作、政治影响以及行为所产生的作用,在项目与不同的委员会、资助人、合作者、竞争者、顾客等之间建立的直接联系。

滚雪球方法(Snowball sampling)能帮助确定出项目所涉及的不同组织和人群。当这些人被确定下来并与之取得联系之后,评估者可以向对项目怀有浓厚兴趣的人群代表进行咨询,从而获得一些有用的信息。这些代表也要回答同样的问题。当这个程序再也不能产生新的重要代表时,评估者便可以确定所有重要的项目方已经被识别出来了。

如果评估是以协作或是参与的形式出现的,这样,某些项目方就会直接参与到设计和评估规划中来(正如第 2 章所叙述的),他们当然就拥有决定评估内容的首要决策权。同样,组织内部的评估者最有可能从项目人力资源那里得到直接的信息。这些项目方的参与使评估逐渐地形成,然而这种方法不能充分地搜集项目各方的观点。或许有些重要的项目群体还会被忽略,而这些人对项目和

评估往往持有相当特殊而且重要的观点。在评估执行程序中,或许这些群体的成员会发表一系列重要的观点,因此,有必要跨越评估团队的指定成员界限去更广泛地征求意见。

所以,总体而言,形成具有可答性的评估问题需要评估小组成员与项目各方成员进行协商。与那些很少甚至没有展示项目各方意见的评估小组相比,那些已采纳了项目各方观点的小组可以相对减少与项目各方的接触。如果评估成员在最初没有与项目各方合作,他们就要重新考虑与重要项目方的交流,从而确保能够完全反映项目各方对评估设计和执行的观点。同样,项目各方的智囊团、执行委员会、或者重要项目方的代表,在与评估者进行正常交流之后,会做出不同的安排。有关这些方法的程序和益处的更多信息能在这些资料中找到,包括费特曼、卡夫特莱恩和旺德曼(Feterman, Kaftarain, and Wandersman, 1996)、格林(Greene, 1988)、马克和萧特兰(Mark and Shotland, 1985)以及巴顿(Patton, 1997)。

在有组织的安排之外,评估者还可以通过访谈的形式,询问项目各方对重要评估问题的看法。通过与这些项目方的早期接触,可以询问有关评估发展方向和调查的问题。这样的访问是典型的无结构或者半结构式的。从项目各方那里调查获得的信息或许也可以从焦点群体中得到(Krueger, 1988)。焦点群体在获取信息方面有特殊的优势,而群体互动在模拟意见及观测值方面有便利效果。但它们同样也有劣势,最显而易见的是,可能会遭遇易变的政治处境和机密性缺失所带来的冲突。在某些情况下,项目方会在与评估者一对一的交谈中,更加坦率地陈述对项目和评估的看法。

对于评估者而言,要从每个项目方的每个成员那里获取信息几乎是不可能的,同样,要完整地识别评估需要涉及的主要事件与问题也是很困难的。通过仔细挑选项目各方的知情者,可以比较顺利地完成上述任务。因为他们与项目有着特殊的关系,因此他们的意见对于确定主要的事件有着特殊的作用。当评估者在与不同项目方的讨论中再也发现不了新的观点时,这便意味着最重要的问题也许已经完全被确定下来。

与项目各方讨论的议题

在需要进行评估时,由评估主办方所确定的问题通常需要与其他项目方进行进一步的讨论,来厘清这些问题对于各方具有怎样的意义,并考察这些问题究竟能够提供哪些有用的信息。在这些讨论中应该涉及哪些问题,在很大程度上取决于评估情境和条件的特定性。我们在此列举几个经常牵涉的议题。

为什么需要做评估? 找到做评估的原因对于评估者而言通常很有意义。评估有可能是被外界需求推动的,在这种情况下,要认识到这种需求的本质和了解评估结果的作用是十分重要的。为了确定项目的有效性,找到改善评估的方法,或者是"证明"项目对潜在投资者、捐赠人和评论家等诸如此类人群的意义,项目经理往往会希望进行一次评估。有时,做评估的推动力量仅仅是政治因素,例如

将有争议性的项目延迟执行。无论原因如何,这些原因都为确定评估应当为谁回答哪些重要问题提供了思路。

项目目的和目标是什么? 毫无疑问,一个项目是否能够达到目的和目标与评估需要了解的重要问题息息相关。了解目的和目标之间的区别至关重要。在项目中,**项目目的**(Program goal)是范围广泛的概念,往往用相当抽象的文字进行陈述。例如,为无家可归人士安排的项目的目的或许是在其城市的覆盖区域中"减少无家可归人士"。尽管这一点很容易被人理解,但是这样的目的太抽象模糊,人们无法确定是否能达成目的。"减少的数量"是5%、10%、还是100%?项目所指的人群仅仅是生活在街巷上的人,还是同样包括只有一个栖身之地或者临时性住处的人?对于评估而言,这些抽象的目的应该被转译成更加细化的陈述,具体陈述所处的环境,以及一个或者几个可测量的"成功"标准。评估者通常将这些可测量成果的具体表述称为**项目目标**(Program objectives)。相关项目目标的集合,表明了要达到项目目的有必要取得哪些特定的成果。专栏3—F列出了确定目标的有益建议。

因此,对于评估者而言,一个重要的任务就是与其他项目方合作,从而确定项目目标,并将具有宽泛、模糊的意义,或是过于理想化的想法转化成为清晰、明了、具体的目标陈述。目标对形势解释越直接、越依赖于可以直接并可靠地观察到的现实,就意味着评估越有意义。进而言之,评估者、评估主办方以及其他项目方应该达成共识,认定项目目标就是评估的核心,这样,项目目标就可用作评估标准,用来判断这些目标是否已达到。例如,如果一个工作培训项目的目标是保持较低的失业率,在评估设计之前,重要的项目方将会同意将这点作为核心问题。

如果没有形成对于重要目标的共识,一种解决方法就是将不同项目方所提出来的意见都包括进来,或者是将那些从相关领域的最新观点和理论中提炼出来的目标包括进来(Chen, 1990)。例如,一个工作培训项目的主办方仅仅只对项目实施后的雇佣频率和时间长短感兴趣。但是评估者或许想知道这些人生活安排的稳定性、处理财务的能力以及获取辅助教育的努力,将这些作为项目产出,因为这些生活方式特征或许也会经历令人乐观的变化,从而促使人们提高就业能力以及与工作相关的技能。

评估中需要回答的最重要问题是什么? 参照巴顿(Patton, 1997)的观点,那些具有优先性的评估问题,应当是最能提供可资利用信息的问题。评估的结果很少会被评估者或评估主办方理解为"为了解而了解"。评估就是要让评估结果成为有用的信息,为项目决策人员所利用,无论是作为日常管理决策的标准,还是更重大意义上的投资或政策管理的标准(见专栏3—G有关评估经理对该程序的观点)。

专栏3—5 确定目标

制订有用的目标有四个技巧:①使用表达力强的动词,②仅仅陈述一个意图或者目标,③确定一个最终产物或者结果,④确定达成预期成果的时间(Kerschner Associates, 1975)。

"表达力强"的动词将会体现一种可观察或测量的行为。例如,"加强健康教育材料的使用",这种动作取向性的叙述包含了能被观察的行为。相反,"将鼓励更多地使用健康教育材料",就是一种表达力相对弱、而且不太精确的陈述。"鼓励"是一个有多种解释的术语。同样以动作为主的动词包括:"写","满足","发现","增加"以及"签署"。表达能力弱且不很精确的动词则有:"了解","激励","提升"和"鼓励"。

获得清晰目标的第二个建议是,仅仅只陈述一个意图或目标。大多数项目都有许多目标,但在每个目标中仅仅只需要陈述一个意图。一个有着两个或是多个意图的目标,将需要不同的完成方法和评估策略,这样将会导致很难达到目标。例如,类似这样的陈述"为孕妇提供三节产前教育课,并为每次课中的25个妇女提供交通服务",达成这样的目标就有困难。这个目标包含了两个任务——提供生育前的教育课程以及提供交通服务。如果只完成了一个任务的话,那么在多大程度上达到了预定目标呢?

确定可用目标的第三种技巧是,确定一个最终产出或者结果。例如,类似这样的陈述"通过与城市医院转交合同,为孕妇提供三节产前教育课",这里包含两个结果——三节课和合同。最好的方法是将这些目标分开陈述,尤其是当这种较高层次的目标(开始三节生育前教育课)的达到部分地取决于一个较低层次目标(签一份转交合同)的完成时。

一个陈述清晰的目标必须只有一个企图和一个结果。例如,"与健康部门建立合作关系"的陈述只表明了目标,但没有表明最终产出或者结果。怎样才能表明他们进行了交流合作呢——电话、会议、还是报告?如果没有将最终产出说明清楚,要进行评估就十分困难。

要写出和评价目标,需要记住两个问题:第一,无论是否有这方面的知识,在人们读了目标陈述之后,是否就能了解到项目的意图;第二,怎样才能判明目标已经达到,并能够展现可见的、可测量的或有形的结果。意图或目标所描述的是要做的事情;最终产出或结果显示的是已做过的事情。这样才能保证"一旦看见,便能了解"。

最后,确定实现目标的时间也很有用。例如"尽快地建起一个诊所"就不是一个有用的目标,因为"尽快"所表达的意思特别含糊。因此,规定一个实现目标的期限是十分有必要的。如果不能确定特别准确的日子,也可以设定一些目标日期段,例如,"在3月1号和3月30日之间的某个时候",这也很有作用。

资料来源:Stephen M. Shortell and William C. Richardson, *Health Program Evaluation*(St. Louis, MO: C. V. Mosby, 1978), pp. 26-27.

专栏3—6 有关评估利用的问题

一位在社会服务组织任职的评估经理总结了他对项目决策者利用评估结果的一些看法:

1. 评估利用或是研究结果的利用并不是孤立的事情。评估报告是枯燥的东西,只有在利用和执行评估结果时,才会引起人们的兴趣、导致人们的行动。因此,评估的利用必须从书面报告变成项目经理的日程安排。

2. 评估利用(在这个过程中项目经验被识别出来)通常要求行为或政策的改变。这需要改变管理人员的操作计划,并对行为的优先秩序进行调整。
3. 评估研究利用也涉及政治活动。这主要取决于人们在组织行动中的实际权力。作为评估的结果,如果要对项目或组织进行改变,需要得到来自最高管理层的支持。
4. 为了使组织安排的程序变得正统且形式化,有必要建立执行评估利用的系统。否则,评估利用很可能会成为个人的事情,这样,评估便会成为某个组织或个人掌握权力和进行控制的工具。

资料来源:Anthony Dibella, "The Research Manager's Role in Encouraging Evaluation Use," *Evaluation Practice*, 1990, 11(2): 119.

不幸的是,评估者在评估中获得的大量经验往往被听取发言的人所忽视。有很多原因促使这种情况发生,其中有许多情况都处在评估者能够控制的范围之外。在评估的最初阶段和完成阶段之间,项目处境可能会发生变化,在获得评估结果之后,评估结果也可能会变得与实际情况不符。由于评估可能不会为决策者提供有用的信息,因此,就会缺乏对评估的利用。在更多的情况下,由于人们对这些工具缺乏认识,这种愚蠢的事情就会在不经意间发生。例如,评估规划看起会获得相关信息,但是,当人们处理这些信息时,却发现其并不像搜集时想象的那样有用。另外,当评估结果所针对的人群在最初对搜集这些信息的目的不是很清楚的时候,这种情况也很容易发生。

在考虑到这些问题之后,我们主张评估者运用回溯性规划来设计评估问题,这种技术先确定所需要达到的结果,然后再回溯先前开始阶段的工作,考虑如何达成这些结果(Elmore, 1980)。采取这种方法时,与评估主办方和重要项目方讨论的核心问题应该是,谁将使用评估结果以及使用这些结果的目的如何。应该指明的是,这个问题所考察的并不是谁对评估结果感兴趣,而是谁将使用这些评估结果。评估者想尽量清楚地了解细节性问题,他们想知道谁会使用评估,以及他们使用评估的目的如何。例如,项目执行者和项目指导者会试图使用评估结果,从而能帮助他们确定下一年所需优先考虑的问题。或者是,当立法委员会忽视了项目的某些领域时,他们希望能获得有关继续为项目提供经费的信息。或者是,发起这些项目的政府机构希望知道,这些项目是否代表了一种成功的模式,并且是否应该被推广到其他地方。

在每种情况下,评估者应该分别与每个评估用户合作,这样他便能够描述用户准备做出的某些潜在决定或是行为范畴,了解用户考虑与此相关问题的形式和本质。为了使这种行为达到最具体的水平,评估者甚至会创造虚拟信息(例如,这样评估的结果将会是,"在 30 天之内,完成项目的顾客中有 20% 回复到以前的情况。"),并与潜在用户讨论这种情况发生会产生什么样的影响,以及他们如何使用这些信息。

对评估结果使用的详细说明以及对有用信息性质的描述,会直接导致评估问题的形成(例如,"在第一个月里,在完成了项目干预的服务对象中,恢复到以

前状态的个体所占比率为多少?"),并有助于为评估问题设定优先级别。在这个结合点上,还必须考虑时机的选择。有些问题必须在其他问题得到了解决之后才能被解答;有时,评估发现的使用者因为其决策进度表,需要在得到某些问题的答案之前得到其他一些问题的解答。于是,最为重要的问题将被结合到不同的分组中去,按照适当的时间顺序安排出来,并与预定使用者的咨询最终完成的形式相结合。有了这些条件,制定评估规划在很大程度上就只是一种追溯性工作,来确定采取怎样的测量、观察和程序等工作,从而在用户确定的时间内,根据其需要,为重要的问题提供答案。

对项目假设和理论的分析

在本章的论述开始时,我们已经提到,除与相关项目方进行商讨之外,评估者通常应当对项目做出独立的分析,来设计恰当并且相关的评估问题。大多数评估问题以"假设发生的情况是否正在发生?""服务客户的目的达到了吗?""服务项目是否充分?"或者"目标达到了吗?"这类主旨为核心表现为不同的形式。那么,一种有用的确定重要评估问题的分析方法是对评估进行细节描述,阐明项目计划实现的目标。评估者可以构建一种概念框架,用来考察项目运作的方式和在不同活动及功能之间应当具有怎样的关联,以及要创造怎样的社会效益。这种表述可以用来确定促使项目发挥绩效的因素。而这些因素能帮助识别重要的项目假设与预期是否适宜,以及项目是否以有效的方式被执行。

我们在这里介绍的是对项目理论的解释,关于项目理论项目已采用的战略和策略之间关系的假设以及项目意图创造出的社会效益。理论这个字眼似乎有些夸张,很少会有项目指导者声称自己正在使用某种特殊的理论。然而,在字典中有关理论的释义是,"完成某些任务所秉承的特殊概念或观点,以及完成这些任务所使用的方法。"通常情况下,当评估者使用项目理论时,他们就是在这个意义上使用理论一词的。或许还有其他一些说法,例如概念或者项目计划、蓝图、设计。

长期以来,评估者已经意识到将项目理论作为评估基础的重要性,这个基础将帮助评估者形成并排序评估问题、设计评估研究以及解释评估结果(Bickman, 1987; Chen and Rossi, 1980; Weiss, 1972; Wholey, 1979)。然而,在具体使用过程中,评估理论却有过许多不同的名字,例如逻辑模式、项目模式、产出草图、因果图、行为理论,等等。在大多数情况下,对于项目理论是否被利用到最佳程度,大家并没有统一的意见,在关于评估的文献中可以找到许多不同的观点,尽管这些观点表现出许多共同的元素。由于项目理论本身就是评估的一个潜在对象,我们稍后将在第5章中全面探讨应当如何表达并且评估项目理论,这对于本议题具有独到的贡献。

我们在这里想要强调的是,理论在形成评估问题的项目分析工作中的工具性作用。举例来说,理论的常见表现形式是逻辑模型,这个模型展现出项目服务结果的预期步骤序列。专栏3—H展示的是未成年妈妈为人父母项目的逻辑模

型。这个项目期望服务机构从当地高中招收怀孕的未成年女性参加为人父母的学习班。这些学习班应当就某些主题展开教学,而未成年妈妈则相应地应该掌握产前营养以及婴儿护理的知识,这些知识又会产生合理的膳食和婴儿护理行为,并最终确保孩子的健康。

通过对项目逻辑的简单罗列,使确定评估应该关注哪些问题变得相对容易。例如,在项目的覆盖范围内有多少符合条件的未成年已怀孕女性,以及她们关于营养和婴儿护理的知识缺乏程度如何?真正参与到这个计划中的未成年已怀孕女性的比例有多大?课程是否囊括了所有应当讲授的主题,这类项目指导的性质是不是适合未成年听众?她们从这些课程中真正学到了什么,最重要的是,作为结果,她们的行为发生了怎样的改变?她们的孩子在十二个月之后是否比没有这个项目存在的情况下更健康?

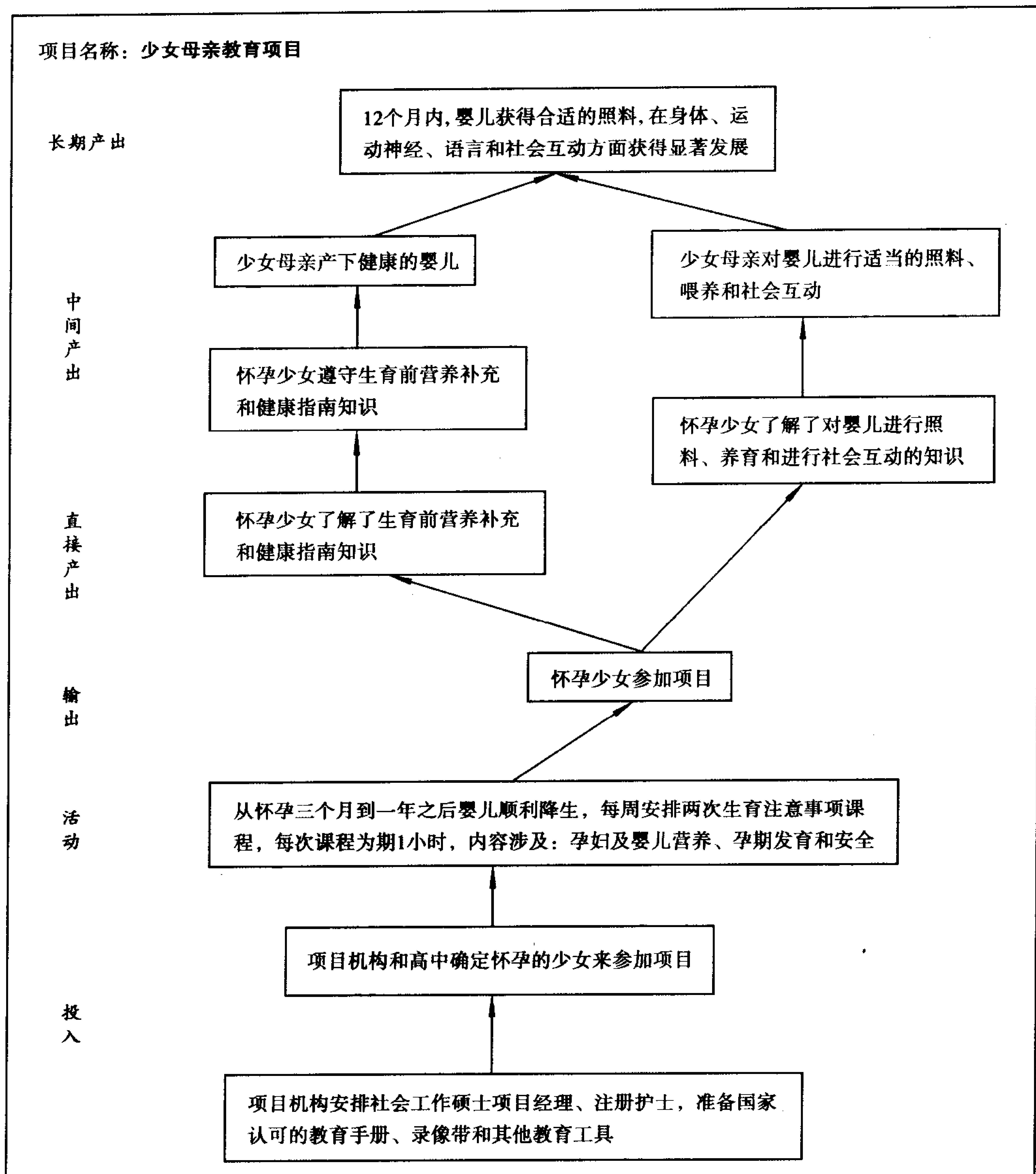
问题中的每一个都可以进一步被提炼为更精细具体的形式,而评估者需要将那些更具体的形式作为其评估设计的基础。在相应领域中,对于项目绩效的预期也可以被进一步地细化为评估绩效的具体标准。如果在与相关项目方的合作中,逻辑模型或者理论图景的其他形式被建构起来,并且代表了各方的共识,那么作为其结果而产生的评估问题基本上是有意义的且具有潜在的重要性。对于评估者而言,特别有益的是这种技术常常能够产生有显著意义的项目议题以及问题,而在不采用这种技术的情况下,这些议题和问题是相关项目方和评估者都认识不到的。

核查评估问题和确定问题的优先秩序

在充分考虑了项目各方所关心的问题和分析了由项目理论指导的项目问题之后,评估者将提出评估会涉及的许多问题。此时的任务就变成针对各个特定主题组织问题并确定它们之间的优先秩序。

问题的组织往往是十分简单的。评估问题围绕不同的项目功能进行聚集(例如,招募、服务、结果),还有,正如前面提到的,也围绕不同的评估事件(需求、设计、执行、冲击、效果)展开。同时,评估问题往往是由许多非常详细的问题组织起来的等级结构(例如,“在公共住房项目中,老年人是否知道这个项目?”),其基础是范围更加广泛的问题(“我们是否接触到了目标人群?”)。

专栏3—11　对少女母亲进行为人父母教育项目的逻辑结构



资料来源：United Way of America Task Force on Impact, *Measuring Program Outcomes: A Practical Approach.* Alexandria, VA: Author, 1996, p. 42. Used by permission, United Way of America.

要从问题中确定重要的成分就更加具有挑战性。一旦对问题进行了勾画，对评估者而言，在规划过程提出的有关项目的大多数问题就会和项目各方的兴趣联系起来。因此，很少有直接针对所有问题的可用资源。这样，对于评估者而言，注重评估的目的和将评估结果合理利用就尤其重要。至于在此过程中投入多少时间和努力，对项目各方而言没有多大意义。

这就是说,我们必须警惕对有用信息的理解过于狭隘。评估利用研究表明,在决策过程中对评估信息操作性和工具性的运用,仅仅是评估信息做出的贡献之一(Leviton and Hughes, 1981; Rich, 1997; Weiss, 1988)。在很多情况下,同样重要的是概念性和延续性的使用——评估发现的贡献,这个项目和社会问题被理解和争论的方式。评估常常确定议题、框定分析,并帮助将注意力集中到重要的问题上来,即使没有直接的证据表明评估发现与某个具体决策有直接联系。这方面议题将在第 12 章中有更为完整的讨论;在这里,我们的目的仅仅是指出,即使是在直接效用或直接使用者不明显的情况下,一些评估问题也是十分重要的。

经过了一些合理的程序,总结了重要的优先评估问题,评估者已经准备好了设计评估的最重要部分。在这本书剩下的部分中,大部分的篇幅将讨论与这个任务相关的方式、方法和需要考虑的问题。这样,接下来的讨论自然就按照评估进展的逻辑进行,组织评估问题和确定重点,进而反过来强调项目需求、项目理论或强调需求的规划、项目的执行计划及其过程、针对社会需求的项目产出或影响以及项目获得预期产出的效率。

小 结

- 评估规划中一个十分重要的方面是对评估将涉及的问题进行确定和阐述。这些问题将评估的注意力集中到主要项目方关注的项目绩效维度上,从而指导设计,以便提供有关项目绩效的有意义的信息。
- 好的评估问题必须是合理并且恰当的,同时也必须是可以回答的。也就是说,好的评估问题应当是能够得到清晰界定的项目绩效的可观察维度,这些可观察维度关乎项目目的,并且代表了项目能实际产生预期成果的领域。
- 使评估问题显得格外突出的,是相关的标准,通过这些标准,就可以判断项目绩效。如果对评估问题的阐述能包含由重要项目方提出的绩效标准,评估规划就会变得更加简单,同时,在对结论进行解释的时候,反对的潜在因素就会减少。
- 评估的议题可以被建构为一个有用的等级结构,其中那些最基础的问题处于底层。通常,处于等级结构较高位置的问题,预设了对较低层级议题的了解或者可靠性的设定。
- 为了确保评估设计中包含了最重要的问题,最好是在与评估主办方和其他项目方以及对项目决策起重要作用的人士进行了交流和协商之后,再将评估问题总结出来。
- 尽管项目各方的意见十分重要,评估者也必须做好确定项目问题的准备,以防万一。这需要评估者对项目假设和预期进行相关的分析。
- 揭示项目绩效的有用方法在评估分析中也很重要,就是将项目理论细致地勾画出来。项目理论描述了项目的假设、项目活动以及这些活动如何能导出项目期望的社会收益。对于项目理论的批判性分析能够使得一些本来可能被忽视的评估问题浮出水面。
- 完成这些不同的程序后,评估问题的雏形便产生了,评估者必须将其组织成一些互相联系的问题,并将项目各方和专家们的意见考虑进来,从中找出最重要的部分。有了这些重要的优先问题之后,评估者便可以设计能被人们回答的评估问题了。

基本概念

覆盖区域(Catchment area):项目服务所覆盖的地理区域。

执行失败(Implemental failure):项目没有充分地按照已有设计所规定的活动来执行,因而没有创造为改善社会环境所需要的条件。另一方面,在创造出来的条件下,项目没有执行服务计划、提供了不充分的服务、提供了错误的服务,或者在目标人群中提供的服务差异太大。

绩效标准(Performance criterion):一种评估标准,通过与之比较,项目的绩效便能被评估出来。

项目目的(Program goal):通常情况下是一句概括性的和抽象的陈述,指明项目发展的方向。可以与项目目标比较。

项目目标(Program objectives):对完成项目所需要达成的一系列具体目标的系统陈述。

理论失败(Theory failure):项目按照计划完成,但是项目服务却并没有立即对参与者产生预期的影响,或者预计的最终社会收益没有达到,或者两者都失败。

4 需求评估

准确描述项目意图消减的社会问题的性质，对于项目评估而言非常重要。解答这些问题的评估活动通常被称为需求评估。从项目评估的角度来看，需求评估是一种手段，借此，评估者确定是否真的存在实施项目的需求，如果确实存在这种需求，那么什么样的项目服务最适合满足这种需求。这样的评估对于新建项目的有效设计是重要的。不仅如此，这种评估对于既有项目同样适合，因为，在很多情况下，不能简单假定项目是必须被实施的或者该项目所提供的服务能很好地契合需求的性质。

当然，对项目而言，需求评估之所以重要，是因为如果在开始时根本不存在什么问题或者项目所提供的服务与问题实际上无关，那么项目就不能有效地解决问题。本章将讨论评估者在诊断社会问题中的角色，讨论评估者如何通过系统的、可重复的程序来完成这一任务，并且，这些程序和方法与社会问题的诊断、干预项目的设计及评估相关联。

正如我们在第1章所描述的，项目评估是改善社会状况的工具。对于突出的社会需求，项目是以似是而非的模糊方式体现，还是针对实际处境进行针对性的回应，对于评估社会项目的绩效而言，这些都是非常重要的问题。

对于给定的、针对社会问题的项目而言，回答这些问题首先是描述社会问题的需要。通过这样的描述，评估者能够了解项目理论是否包括了问题的概念化以及修正问题的适当方法。如果回答是肯定的，那么关注点则转向项目的实施是否能与理论相一致，如果是的话，那么对社会状况的改善是否真能实现以及需要付出多大的代价。这样，在对项目期望改善的社会问题的小心描述中，就建立了项目评估的逻辑。

那些被评估者和其他社会研究者用来系统地描述、诊断社会需求的程序，通常被称为**需求评估**。

作为需求识别者的项目评估者，其基本任务就是以尽可能谨慎、客观和长远的方式描述项目各方所关心的"问题"，并帮助诊断以建构有效的干预。这一任务包括：界定社会问题的精确概念，评估社会问题存在的范围，定义和识别干预对象，准确描述干预对象所需服务的性质。这一章将详细讨论以上这些内容。

评估者在诊断社会状况和服务需求中的角色

在事物的宏大系统中，与政党、赞助团体、记者和各种具有领袖气质的名人的较为重大的行动相比，评估者对于识别和缓解社会问题作出的贡献则显得朴实。关注社会问题的动因主要来自政治或道德领袖和社会活动的倡导者，他们由于自身或职业的条件，有能力去处理具体状况。譬如，在二战后对精神疾病的关注就深受一位单身国会议员的影响；为智障人士设立的联邦项目在肯尼迪(Kennedy)总统期间有了很大的增长，因为他有一位有智障的兄弟；机动车安全措施的提高很大程度上归功于纳德(Nader)的倡导；因为媒体曝光和利益与压力团体活动，以及自身有需求的人们有组织的努力，控制不合法和不合格医疗与福利服务的努力时常涌现。

尽管如此，评估者对于改善人类和社会状况仍然做出了巨大的贡献，但并不是通过动员或疏解社会不满群体、摧毁社会中种种阻碍性因素以及在社会的边缘发难的方式。相反，他们以一种平凡却很重要的方式，通过将他们的研究技术应用于系统地描述社会问题的性质，通过评估计划将要实施的干预或已经实施中的干预项目的适当性以及这些项目对社会状况改善的有效性来发挥作用。

作为结果诊断信息的重要性不能被夸大，因为推测、印象式的观察判断、政治压力，甚至偏向性的报告，都会激起政策制定者、设计者和资助机构采取行动，支持或停止支持正在进行的项目。但是，如果要得到合理评判的话，全面认识和理解项目提出的社会问题的性质和范围，以及相应的项目对象和项目实施或将要实施的环境非常重要。这里有几个案例，说明忽视充分诊断程序所可能产生的后果。

- 内陆城市的高失业率问题经常被认为是周围地区就业机会短缺的反应,由此设立了能够对内陆城市工商企业带来大量刺激的项目。从后来的经历中发现,这些工商企业所雇的大部分工人来自于被认为是需要帮助的周边地区。
- 许多1960年代城市翻新项目的设计者认为,那些住在被设计者认为是破旧房屋的人们同样也认为他们的房子是有缺憾的,因此支持拆迁计划。但是在其他城市,位于城市翻新改造地区的居民却强烈地反对这些拆迁计划。
- 旨在鼓励人们进行体检以发现癌症早期迹象的媒介项目,使得医疗中心陷入病人络绎不绝的困境。媒介的努力使得许多没有癌症病的忧郁症患者相信自己得了癌症。
- 在提高对艾滋病临床识别的努力中,包括了给社区医生讲授通过使用血检诊断高危病人综合症的细节。这些知识传播下去以后,发现很少有医生继续把询问性生活历史当作日常诊断内容之一,进而他们不再愿意知道病人中谁最容易被感染。取而代之的是,运用新学来的知识,对所有病人进行血检,结果是要进行大量的检验,花费很高的代价,并给病人带来风险。
- 通过生育控制计划的推广,来减少某大城市中心区的高堕胎率,但是此计划并没有吸引到更多的参与者。随后发现许多有此意向的对象已经得到很好的服务,而且有很大比例的人采用避孕法。高堕胎率主要是由专程为堕胎而从农村来到城市的年轻女士造成的。
- 持枪犯罪问题引起立法提案禁止与犯有重大案件的人进行枪支交易。但是,多数罪犯并不从合法枪支贩卖商那里购得他们的枪支,而贩卖商也没有可靠的方法来辨认购买者是否有犯罪记录。

在所有这些例子中,好的需求评估能够提供信息,进而对问题做出有效描述,从而防止项目提供不适当的或不需要的服务。在一些情况下,因为问题并不存在,所以设计出不必要的项目。在其他情况下,或因为目标人群不需要项目提供的服务,或项目干预被误解,或不可能甚或未能按项目计划实施,进而使得项目不能按预期发挥作用。

所有社会项目都立足于一组前提假设:对项目问题性质和服务的目标人群特点、需求及反馈的描述。任何针对新项目规划、既有项目的变动或者正在实施中项目的有效性的评估,都必须结合这些前提和描述。当然,对问题的诊断和对目标人群的描述已经很好地、令人信服地建立起来之后,评估者就可以把这些作为已知信息,并进入下一步的分析。也有这样的情形,即项目的需求和需求的性质并不需要进行单独考察。事实上,项目的人事部门和资助者经常认为,他们对社会问题和目标人群需求已经足够了解,因此,任何进一步的调查都是浪费时间。不过,这种情况必须慎重对待。就像上面的例子,无论是由于最初对问题的诊断不充分,还是在项目开始后问题和目标人群发生了变动,又或者由于选择或

刻板印象导致了偏差，都容易使项目建立在错误的假设基础上。

所以，在各个方面，评估者都应该详细研究关于项目问题和目标人群的假设，这些假设体现了项目的性质。不管哪里出现模棱两可的问题，评估者都要与主要项目方一起，系统而明确地重新表述这些假设，从而把这些假设作为项目评估设计和评估理论的充分标准。同样，这些假设对评估者实施一些细小的、涉及项目目标与目标人群的独立调查，也是有用的。对于新项目或者其绩效已被纳入讨论的既有项目，对项目服务的目标人群的社会需求进行彻底评估是有必要的。

值得一提的是，需求评估并不总是为某个具体社会项目和项目计划提供参考，需求评估的技术同样可以用来作为决策者的规划决策辅助工具，这些决策者必须在众多竞争性的需求和申请中考虑优先秩序。例如，某区域联合体或某大都市市议会或许委托一项需求评估来帮助他们决定怎样通过各种服务领域来分配资源。或者某个州的精神健康部门可能通过对不同精神健康服务机构的需求评估来在各服务机构间进行最佳资源配置。因此，需求评估通常还包括：根据问题的严重程度、被忽略程度和显著程度，进行需求排序。尽管对某个具体项目（无论是既有的还是被提议的）的需求评估在范围和目的上有所不同，但一般的需求评估所使用的方法却非常相似，而且通常由评估研究者来操作使用。

专栏 4—A 中提供了一项关于需求评估基本步骤的概览，这一专栏强调在需求评估的各个步骤中，项目各方的参与和卷入，这些内容在其他章节还会有大量的讨论，本章也就没有必要进行详细阐述（关于需求评估的应用与技术的长篇讨论，参见 McKillip, 1987; Reviere et al., 1996; Soriano, 1995; Witkin and Altschuld, 1995）。

专栏 4—A 需求分析的步骤

1. 使用者与使用的识别。评估的使用者是那些以结论为基础办事的人和那些受分析影响的对象。这两组人的卷入通常为分析和建议的实施提供了便利。知晓需求分析的效用能够帮助研究者集中关注他们所思考的问题和解决方案，但同时也可能妨碍步骤 3 提到的问题与解决方案的识别。
2. 目标人群和服务环境的描述。地理分布、交通、目标人群的人口统计特征（包括承受力）、资格限定和提供服务的能力都是重要的。各种社会指标经常用来直接或经推算间接描述目标人群。详细记录可以利用的服务资源清单，能够发现在各种服务与补充的以及竞争的项目之间的差距。比较那些享用各种服务的人群与目标人群，能够发现目标人群中未满足的需求和阻碍解决方案实施的障碍。
3. 需求识别。这里描述的是目标人群中存在的问题和可能的解决方案。通常会应用多种信息源。识别中应包括以下信息：对结果的预期，目前的产出，解决方案的效力、可行性，解决方法的利用。通常使用多种方法，如各种社会指标、调查、公共论坛和直接观察等。
4. 需求评估。一旦问题和解决方案被确认，就需要将其综合起来分析以提出建议、指导行动。在过程中可运用定量与定性的综合方法。这个过程越明确、越公开，结论被接受和实施的可能性就越大。

5. 交换意见。最后，需求分析的结论必须在决策者、使用者和其他相关的对象之间交流。为这一步骤所花费的努力应与需求分析的其他步骤相对称。

资料来源：Jack McKillip, "Need Analysis: Process and Techniques," in *Handbook of Applied Social Research Methods*, eds. L. Bickman and D. J. Rog (Thousand Oaks, CA: Sage, 1998), pp. 261-284.

界定社会问题

关于政策变动的建议，新的或修改过的项目，或者对既有项目的评估，通常都源于一个或多个资助方对现有政策和项目绩效不满或者发现新的社会问题正在出现。无论是哪种情况，一个社会问题一旦被确认和定义，那么，事情就不会像看上去那样简单直接。事实上，如何界定社会问题已经让精神领袖、哲学家、社会科学家思索了几个世纪。这一过程比较棘手的问题是，针对干预对象需要或期望决定项目需求的意涵，并决定使用什么标准来识别与项目目标或预期相关的需求。对于我们的目标而言，关键在于社会问题本身不是客观现象，而是与所观察环境相关且卷入政党利益所形成的社会建构。在这种意义上，社区成员以及卷入特定争论的项目各方逐步建构出一个公认的社会问题的社会事实（Miller and Holstein, 1993; Spector and Kitsuse, 1977）。

例如，贫困被公认为是一种社会问题。可观察到的事实是关于收入和财产分布的统计资料。然而，这些统计资料并没有界定贫困的概念，而只能根据已给出的定义来决定多少人处于贫困中，统计资料不能把贫困确立为一个社会问题，而只能描述一种被个体和社会机构视为有问题的社会状况。而且，无论是贫困的定义还是改善贫困者状况的项目对象，在一段时期内，在不同社区和项目各方之间，都是变化的。所以，为减少贫困所采取的主动行为就从增加就业机会和减少经济流动障碍到仅仅是降低低收入者期望值的范围不一。

这样，界定一个社会问题和明确说明干预对象基本上是一个政治过程，而不能仅仅依据问题的内在特性来定性。这种情况已被很好地例证，例如由美国审计总署进行的一项关于减少青春期怀孕的法规分析（GAO, 1986）。美国审计总署发现：没有任何一个即将出台的立法提案把孩子的父亲纳入到讨论的问题中，每一个提案都是把青春期怀孕看作是母亲的事。尽管这种看问题的视角或许会引出了有效的项目，但是很显然，还存在把父亲包括进来的替代概念。

事实上，一个问题的社会界定已经成为政治回应中很重要的一部分，以至于法规的导言经常不辞辛苦地详述该种提案所要补救的社会状况。例如，两个争论中的立法提案都针对无家可归者问题，但是一个提案把无家可归者看做是没有亲属可以依靠的、非常贫穷的人，而另一个提案则把无家可归者界定为无法获得正常的安身立命之所的人。第一个定义主要侧重潜在群体的社会疏离上，而第二个则集中在住宅供给的安排上。根据这些定义而采取的合理改善举动也会

不尽相同,例如第一个定义将支持试图调和无家可归者与其疏离亲属之间关系的项目,而第二个则拨款资助住宅供给项目。

因此,确定一个项目提出的问题在特定政治环境下的意涵,对于评估者来说通常是有益的。例如,为了调查项目提出的问题,评估者会研究出现在政策和项目建议中的各种隐含的或明确的定义。启发性的信息也可以在立法会议录中获得,包括委员听证会、议员辩论、期刊观点、报纸和杂志的评论以及其他讨论该问题的渠道。一个具体项目问题的操作化定义通常可以在项目文件、项目开始的新闻报导、投资计划书和其他诸如此类的地方看到。这些资料或许明确地表述了问题的本质和项目所要实施的计划,如在投资计划书里;或许通过假设(对于项目活动、成功与否以及计划的表述的基础)含蓄地界定问题。

这样的调查很可能为对项目大致所要反应的社会需求的初步描述提供很有用的信息。照此,还能够引导一个更具探索性的关于问题如何被界定以及什么样的替代观点更适合的需求评估。

同样,评估者在这一阶段所扮演的一个重要角色是提供决策者和项目管理者关于政策和项目中问题定义的批评,并提出或许更可用可选定义。例如,评估者指出,把青春期怀孕问题基本上定义为一种非法生育就会忽略在已婚青年中大量的合法生育,因此建议在项目定义后面加上解释说明。

将问题具体化:时间、地点和范围

如果已经清楚地界定了项目意欲介入的问题,评估者就可以估计相关问题的严重程度。社会项目的设计和经费应当与其所回应问题的广度、分布、密度相一致。例如,在为某社区的无家可归者提供紧急庇护所援助的评估中,无家可归者的总人数是350人还是3 500人将有很大的区别。同样,问题主要是出在贫民区还是在富人区以及有多少无家可归者患有心理疾病、慢性酒精中毒和身体残疾也会造成很大的区别。

确定问题的存在比正确估计问题的密度和分布要容易得多。确证一群经常被打的孩子能够足以让怀疑存在儿童虐待问题的人信服。但是要详析问题的大小以及在地理意义和社会意义上的分布,则需要获得受虐儿童的数量、犯罪人特征、问题分布等详细资料。对于像儿童虐待这样的并非普遍的公众行为问题,要获得以上的详细资料是困难的。这样的社会问题多数是"看不见的",因此只有可能获得发生率的不确切估计。在这种情况下,经常要运用多渠道的资料和使用不同的估计方法(e. g. ,Ards, 1989)。

要估计事件发生率,至少要获得一些相当有代表性的样本。当需要对总体进行估计以判定问题的严重程度时,从比整个人群具有更严重问题的高危人群中进行估计,更容易造成误解。例如,对住在庇护所里的女性群体进行孕期配偶施虐事件发生率的估计,就会导致对孕期妇女受虐待总体发生率的偏

高估计。在更具代表性的样本中获得的估计虽然同样指出了怀孕妇女遭受殴打是一个很严重的问题,但不同的是,这样的估计把问题的程度放到了现实角度去考虑(参见专栏4—B)。

专栏4—B 孕妇遭受家庭暴力的频率估计

所有的妇女都有遭受殴打的危险,而怀孕就更增加了一个妇女遭受严重伤害和不利健康结果(不仅对她自己,也对她未出生的婴儿)的危险性。区域性探索研究发现,有40%~60%的被殴妇女在怀孕期间遭受过虐待。例如,在达拉斯庇护所的542名妇女中,有42%的人在怀孕期间遭到殴打。多数妇女认为在怀孕期和孩子的幼儿期家庭暴力问题尤为严重。在另一项研究中,访问全美270名被殴打的妇女,发现有44%的人在怀孕期间被虐待。

但是多数关于孕期殴打现象的报告是从被殴妇女的样本中,特别是在庇护所的妇女中取得的。为了掌握在代表性产科人口中怀孕期间被殴打情况的流行态势,麦克法兰(McFarlane)及其助手在人口超过300万的大城市的公立或私人诊所中,随机抽样并访问了290名健康怀孕妇女。这290名黑人、白人、拉丁人妇女的年龄分布从18岁到43岁,多数已经结婚,80%的人已经有至少5个月的孕期。这些妇女被询问了9个与虐待相关的问题,例如,她们是否与在目前怀孕期间对她们实施过打、掴、踢或其他肉体伤害的男性伴侣有联系?如果有,受虐待的可能性就增加。在290名妇女中,8%的人在孕期遭到过殴打(每12个被访者中就有1个)。另外15%的人指出在怀孕前曾遭到过殴打。关于殴打的频率并不像人口统计变量函数那样变化。

资料来源:J. McFarlane,"Battering During Pregnancy:Tip of an Iceberg Revealed." *Women and Health*, 1989,15(3):69-84.

运用现有的资料来源做估计

对于某些社会问题,既有研究和资料来源,比如调查和人口普查数据,对于评估某个社会问题的特定方面,将能够提供高质量的有用信息。例如,美国年度人口抽样调查和十年一次的人口普查所获取的人口资料,就能够为某些问题的研究提供有价值的、精确的信息。美国十年一次的人口普查包含的区域(每个区域包括大约4 000户家庭)调查资料,就能够进一步扩展为评估研究所需的地区或者社区资料。例如,专栏4—C就描述了利用重要的统计记录和人口普查资料来评估佛罗里达州可怜的出生状况问题的本质及其重要性。这项需求评估旨在评估孩子和母亲的健康需求以便设计出合适的服务计划。即使有关感兴趣问题的直接信息不能从已有的记录中获得,如果知道可利用的资料与问题的各项指标之间的经验性关系,也可以做出间接的估计(e. g. Ciarlo, et al. , 1992)。例如,某个区域内的学生所享受到免费午餐的比例,可以作为这个地区贫困水平测定的一个指标。

如果信息来源的有效性并不像人口普查资料那样被广泛认可,那么,在使用这些来源时,有必要仔细检查资料是如何被搜集起来的。如果单凭经验来进行推测,在任何问题上,不同的资料来源可能会提供出截然不同的甚至相反的估计。

专栏4—C 运用重要统计与人口普查资料来评估儿童和母亲的健康需求

在佛州，一项"健康起步"草案（The Healthy Start Initiative）（一系列旨在改善州内怀孕和出生状况的法律措施）为社区胎儿和婴儿医疗看护联盟的建立制订了规则和条款，这个联盟由以下几个单位组成：医疗看护提供者，州或地方政策代表，社区联盟，母亲和儿童卫生组织，计划生育的消费者，胎儿看护和主要的看护服务机构。每个单位都需要对各自的服务范围进行需求评估并要求制订出服务提供计划。关于加斯登居民健康儿童有限公司（The Gadsden Citizen for Healthy Babies, Inc）（代表佛罗里达北部一个基本上是乡村的、有多数非洲裔美国人居住的小县）的需求评估，使用现有资料，取自于佛州重要的统计资料和美国人口与住房（Population and Housing）机构的普查资料，来估计该县的母亲和儿童健康问题的程度和分布。

首先，通过运用佛罗里达重要统计资料（记录州内每年的出生和死亡信息）来调查该州怀孕结构和相关的母亲特征。特别是，检查了以下几个指标：

1. 婴儿死亡率：该县的比率大大高于全国或州的水平。
2. 胎儿死亡率：这个比率高于州的目标，且非洲裔母亲的比率高于白人母亲。
3. 初生儿死亡率：这个比率高于州制订的相对于白人的目标，但低于相对于非洲裔母亲的指标。
4. 初生婴儿死亡率：这个比率低于州目标。
5. 低出生率：对于青少年和年龄在35岁以上的妇女，发生的可能性更高。
6. 初生婴儿低重率：总体比率是州的2倍，且不管是非洲裔母亲的比率还是白人母亲的比率都超过州的目标。
7. 青春期怀孕：在十几岁年龄段内生养婴儿的比率是州平均数的2倍以上；在同一年龄段内，非洲裔青少年怀孕的比率是白人比率的2倍以上。
8. 母亲年龄：在年龄是16—18岁的母亲中，婴儿死亡率和低出生率都是最高的。
9. 母亲的受教育程度：低于中学教育水平的母亲生养低重婴儿的可能性比其他人稍高，孕检中发现，其生养有问题儿童的可能性几乎是其他人的8倍。

基于以上发现，识别出导致不良出生结果可能性高的三个组：

1. 低于19岁的母亲
2. 低于中学教育程度的母亲
3. 非洲裔母亲

从人口普查办公室（Bureau of Census）获得的人口普查资料中，可以找出以上三组中每组育龄妇女的数量、她们在不同低收入阶层的比例，并依据人口普查资料的地域编码和缩位编码，获知她们在该县的地理位置分布。这些资料被项目组织用来识别该县的主要问题区域、确立目标，并制订服务计划。

资料来源：E. Walter Terrie, "Assessing Child and Maternal Health: The First Step in the Design of Community-Based Interventions," in *Need Assessment: A Creative and Practical Guide for Social Scientists*, eds. R. Reviere, S, Berkowitz, C. C. Carter, and C. G. Ferguson (Washington, DC: Taylor & Francis, 1996), pp. 121-146.

运用社会指标来确认趋势

对于一些课题，现有的资料来源能够提供定期的测量以用来绘制历史发展

趋势的图表。例如,人口普查办公室的当前人口调查,每年通过选取大家庭样本来搜集关于美国人口特征的资料。资料包括关于家庭构成,个人收入和家庭收入,家庭成员的年龄、性别、种族等测量数据。收入和项目参与的定期调查提供了美国人口参加各种社会项目的程度的资料,这些项目包括:失业津贴、对有儿童家庭的资助、食物券、职业培训项目,等等。

这些定期的测量被称为**社会指标**(Social indicator),能够从多个方面为评估社会问题和需求提供重要的信息。首先,对于一段时间内追踪其发生发展过程的社会项目,如果分析得当,资料能够被用来评估这一社会问题的大小程度和分布;其次,已出现的趋势能够用来提醒决策者注意某一社会状况是在改善、保持原样还是在恶化。最后,社会指标的指向能够提供初步的或者是粗略的、关于已经就位社会项目的绩效。例如,收入和项目参与调查能够评估全国性项目如发给失业者或贫民的食物券或职业培训所涉及的范围。

社会指标数据经常被用来督导社会项目影响之下的社会状况的变化。目前,就有研究者花费大量的精力来搜集贫困家庭的社会指标数据,以试图判断1996年个人责任与工作机会协调法案中在福利方面实施激进改革之后,贫困家庭生活环境的变化,是恶化了还是改善了。关注儿童健康的专门调查由城市研究所和人力发展研究机构实施进行。另外,人口普查办公室把收入与项目参与调查的范围扩展,建立家庭座谈小组,从而能够在福利改革实施前后进行多次访问(Rossi,2001)。

不幸的是,目前可获得的社会指标受到社会问题研究范围的限制,主要集中在贫困和就业、非法欺诈行为、国家项目的参与和家庭构成等问题上。对于许多社会问题,没有社会指标以资参考,或者虽然有支持国家层次的趋势分析但却不能分解,从而也不能提供对地方层次趋势分析有用的指标。

通过社会研究来估计问题参数

在许多情况下,现存资料来源没有提供某问题的程度和分类估计。例如,尚缺乏关于家庭杀虫剂滥用情况的现成资料,这就导致了这样的质疑:滥用杀虫剂是否是一个问题,例如在有孩子的家庭中。在其他情况下,可在国家的或者整个区域层次样本中利用的有关问题的详细信息却不能分解到相应的地方层次上运用。例如,一个关于家庭药物滥用的全国性调查,采用全国性的代表性样本来追踪药物滥用的特征和程度;然而大部分州的回答者人数不足够大,以至于不能提供在州级层次的关于药物滥用的很好的估计,因而城市水平上的有效估计也根本无法获得。

如果有关资料不存在或者不充分,那么研究者就必须考虑搜集新的资料。从增加投入程度到依靠“专家”指导做大规模的**抽样调查**(Sample survey),对社会问题的程度和种类做评估有多种方法。是否要开展某项研究部分取决于可获得的经费,部分取决于做出精确估计的重要程度。如果对某个政治管区内营养失调婴儿数量的了解是至关重要的,那么在立法或者项目设计时,仔细设计的健

康调查则必不可少。相比之下，如果只是需要确定在婴儿中是否存在营养失调现象，从知识渊博的专家那里得到的信息就已经足够了。这部分简单论及了评估者可以发掘相应资料的三种不同来源。

机构记录

为项目对象提供服务的组织记录是对评估某个社会问题的范围和程度有用的信息来源(Hatry, 1994)。一些机构对客户做了很好的记录，但另一些没能做出高质量的记录或者根本就没有记录。当某个机构的客户包括了所有反映项目对象，并且所做的记录均可靠的时候，评估者就不需要再做任何进一步的调查了。不幸的是，这种情况并不常出现。例如，通过有根据的推断(从药物滥用诊所的病人记录中获得)来试图估计药物滥用程度是一种很诱人的做法。如果所有滥用药物的社区完全被现有的诊所覆盖，这样的估计将很准确。如果药物滥用诊所确实覆盖了所有或者大部分滥用药物的人口，治理药物滥用的项目或许就不会有问题。所以，如果问题被现有项目充分地处理的话，这些项目所得出的数据将是有用而准确的，但是这种情况很少出现。当然，在药物滥用诊所的例子中，所有药物滥用者实际上是否都在诊所得到治疗是令人怀疑的。(在专栏 4—B 关于受虐孕妇例子中，阐明了通过某个项目目标人群和某个总体的抽样样本调查所获得的不同估计)。

调查与人口普查

当有必要获得非常准确的关于问题严重程度和分类的信息而又没有现有的可靠数据可资利用时，评估者需要运用抽样调查或者普查的方法开展原始数据调查。因为任何一项这样的技术都牵扯大量的精力和技术手段，再加上资源的限制，这些调查常遇到不同规模和程度的技术复杂性问题。

举一个极端的例子，专栏 4—D 描述了一项旨在评估芝加哥无家可归人口规模和构成的需求评估调查。这项调查的对象包括了住在紧急庇护所的和没有住在紧急庇护所的无家可归者。对后者的调查内容还包括在午夜对芝加哥各个街道的搜查。因为约翰逊(Johnson)基金会和教会基金组织(Pew Memorial Trust)正要计划一个项目以增加无家可归者获得医疗照顾的机会，所以才有了这项调查。尽管有大量的迹象表明在城市中心的无家可归人口中存在着严重的医疗问题，但实际上既无关于无家可归人口规模、也无关于这些人口中存在医疗问题严重程度的准确和可靠的资料，所以，基金会希望资助一项研究计划以搜集缺失的信息。

然而在通常情况下，需求评估调查并不像专栏 4—D 中所描述的那样精细复杂。在许多情况下，常规的抽样调查就能够提供足够的信息。例如，如果要获得有关需要被看护孩子的数量和分布状态的可靠资料以便计划提供新的相应设施，那么通过电话进行抽样调查来获得这些资料是切实可行的。例如，专栏 4—E 描述了在洛杉矶对 1 100 多人进行的电话调查，目的是探知公众对艾滋病预防绩

效了解的程度。对于旨在提高对艾滋病各种预防方法认识的大众传媒教育项目,这样的调查确认了在公众认识中存在差距的程度和特点。

许多调查组织有能力计划、执行并分析为需求评估而做的抽样调查。另外,还可以对定期组织的研究加入新的提问,借助这些研究,许多组织就可以节省研究时间,从而降低成本。无论使用什么样的方法,都必须认识到设计和实施抽样调查是一个需要十分具体技能的复杂过程(关于对抽样调查方法论各个方面的讨论,参见 Fowler, 1993; Henry, 1990; Rossi, Wright, and Anderson, 1983; Sudman and Bradburn, 1982)。

专栏 4—D 运用抽样调查研究芝加哥的无家可归者

多数抽样调查都基于这样一个假设:所有的对象都能够被统计到,并且在他们的住所被调查到。但在任何一项关于无家可归者的研究中,却不能做出如此假设。所以,为芝加哥无家可归者研究设计的策略就不同于传统的调查,因为样本来自居无住所的群体,而且访问也是在有住所的人和居无定所的人之间的区分达到最大限度时进行。并采用了两个互补性的样本:①在为无家可归者提供的庇护所里过夜的人;②从午夜 12 点到凌晨 6 点在无住宅区(从芝加哥进行人口普查的街区内选取的概率样本)遇到的人的总计数。庇护所调查和街道调查合在一起,构成了关于芝加哥无家可归者的无偏样本。

如果一个人住在为无家可归者提供的庇护所里,或者他在街道调查中被调查者遇到并被发现没有租、拥有房子,或者不是某类家庭——这类家庭或租或拥有一套普通住房——的成员时,这个人就被划分为无家可归者。普通的住房包括:公寓、房屋、饭店或其他设施里的房间,以及活动房屋。

在街道调查中,访问组在芝加哥警察的陪同下,搜索抽样街区内所有可以进入的地方,包括夜市区、小巷、过道、屋顶和地下室、废弃的楼房、停放的汽车和卡车等。街道调查中所有遇到的人,如果有必要,可以把他们叫醒并询问以确定他们是否是无家可归者。在庇护所的样本中,所有在那里过夜的人均被认为是无家可归者。一旦辨认出无家可归者,调查者就需要对他们进行访问以获得有关他们职业、居住历史以及社会人口特征的资料。所有被访者完成访谈后可得到 5 美元的报酬。

资料来源:P. H. Rossi, *Down and Out in America: The Origins of Homelessness* (Chicago: University of Chicago Press, 1989).

专栏 4—E 评估对 HIV 预防知识的了解程度

为了估计人们对如何避免艾滋病病毒(HIV)感染知识的知晓状况,研究者对洛杉矶县居民的抽样样本进行了电话调查。被调查者要求评价四种方法(一些用来避免通过性行为而感染艾滋病的方法)的有效性(见下表)。在他们的评价中,绩效最好的方法是在 HIV 呈阴性的群体中实行一夫一妻制的性生活,虽然有 12% 人认为即使在这种情况下,也不能确保安全;安全套的使用,尽管有破裂、泄漏和滥用的问题,但仍有 42% 的回答者认为是很有效的方法,另外有 50% 的人认为,在某种程度上这种方法是有效的;回答者都不肯定杀精子剂这种方法的绩效,不论这种方法是单独使用,还是与其他方法合用。

不同预防方法绩效的等级分布百分比

预防方法	很有效	有点效	无效	不知道
在 HIV 呈阴性的个体中实行一夫一妻制的性生活	73	14	12	1
只使用安全套	42	50	7	1
使用带杀精子剂的隔膜	9	35	50	6
只使用杀精子剂	7	32	53	8

资料来源：D. E. Kanouse et al. *AIDS-Related Knowledge, Attitudes, Beliefs, and Behaviors in Los Angeles County* R-4054-LACH (Santa Monica, CA: RAND, 1991).

主要知情者的调查

或许，对社会问题严重和重要程度进行估计的最简单方法是询问**主要知情者**(key informants)。这些人的立场和经历使得他们对于问题的重要性和分布状态有一定的看法。不幸的是，这样的报告通常并不准确，这些主要知情者中很少有人能够提出论据充分的观点或信息，进而对受某种社会条件影响的人口数量和这些人在人口统计及地理分布上的状态做出好的估计。例如，评估某个社区内无家可归者的人数。尽管合适的知情者能够提供这类人群的某些资料，但是却很难估计这部分人口的规模。事实上，有资料表明，知情者对他们居住区内无家可归者人数的估计往往范围偏大且倾向于高估，有时还是相当大的高估(参见专栏 4—F)。

考虑到主要知情者提供的关于问题程度的资料总比没有任何资料强，评估者在没有进行其他研究的可能或者可得到的资金不足以支持一个更好的方法的使用的时候，可以组织主要知情者调查。在这种情况下，评估者必须仔细检查以确保主要知情者调查是高质量的。研究者应十分仔细地选取被调查的对象，力争他们掌握必要的专门知识，并确保以小心谨慎的方式询问他们问题(Averch, 1994)。

专栏 4—F 使用主要知情者资料对无家可归者人数的估计

为了探知"专家"对洛杉矶商业区无家可归者人数的估计与无家可归者(在街上、在庇护所、在贫民窟里)的实际人数的相近程度，一组研究者让 8 位在贫民区工作的服务提供者(庇护所的工作人员、社会机构的官员和其他类似的人)来估计 50 个街区的无家可归者总人口。

- 提供者 1：6 000 ~ 10 000
- 提供者 2：200 000
- 提供者 3：30 000
- 提供者 4：10 000
- 提供者 5：10 000
- 提供者 6：2 000 ~ 15 000
- 提供者 7：8 000 ~ 10 000
- 提供者 8：25 000

很显然,这些估计值差距悬殊。其中两名提供者(4 和 5)的估计与研究者估计的最可能的人数——基于在庇护所、贫民窟、街道统计的人数——相当接近。

资料来源:Hamilton, Rabinowitz, and Alschuler, Inc., *The Changing Face of Misery: Los Angeles' Skid Row Area in Transition—Housing and Social Services Needs of Central City East* (Los Angeles: Community Redevelopment Agency, July 1987).

需求预测

在系统地阐述政策和项目并对其进行评估时,通常有必要对某个社会问题以后的重要性做出估计。一个现在看来比较严重的问题,在几年后将会变得更严重或者变得不太严重,进行项目设计时必须把这种趋势考虑进去。然而,预测未来的趋势是十分冒险的事,特别是当时间的跨度增大时就更是如此了。

在预测上存在有许多技术性和实践性的困难,这些预测工作的基础部分一定程度上基于如下假设:认为未来与过去、现在相关。例如,根据初步印象,借助现有的资料推算从现在起到十年后人口中年龄在 18 到 30 岁之间的人的数量似乎很容易做到(几乎完全由现在的人口年龄结构决定)。然而,二十年前人口统计学家做的一项关于中非地区的人口预测,却极大地偏离了实际值,这是由于无法预期的和不幸在青壮年中甚为流行的艾滋病的影响所致。对于时间跨度较长的推算就更可能有问题,因为必须把人口出生率和死亡率的发展趋势也考虑进去。

我们并不是反对在需求评估中运用预测。相反,我们只是提出警告,反对那些未对预测得出的过程进行彻底检查就不加批判地接受预测结果的做法。此外,这样批判性的检查或许本身就存在一些困难。对目前趋势的简单推断,预测所依据的假设相对较少而且容易确定。然而,纵使假设已知,也不易判定是否合理。为了精密的推算,例如那些从多元方程和统计模型中发展出来的推算,假设检验需要高级程序员的技能和富有经验的统计人员的专业知识支持。在任何情况下,评估者都必须认识到几乎最简单的预测也需要专业知识和程序性技术活动的支持。

界定和识别干预对象

从项目各方最早开始对社会问题下定义到项目运行的整个时期,正确界定和确认干预对象对于社会项目的成功至关重要,是探明项目需求的先决条件。然而,一个项目要想更加有效,就必须不仅仅清楚它的目标人群,而且要能够将项目服务直接提供给干预对象,而不是其他类型的人群。相应地,将服务提供给干预对象,需要准确地定义干预对象,从而确保以相对清晰和有效的方法把干预对象与其他非目标人群区别开来,这是项目操作过程中的一个必要部分。

因为关于问题的定义和相应的人口规模估计，随着时间的推进可能会有所变化，因而使得详细描述干预对象变得复杂。例如，20 世纪 80 年代早期，无家可归者问题逐渐突出起来。最初，无家可归者被看做是那些住在街道、小巷或者自己搭建的棚屋里的人。然而，当改善无家可归者状况的倡导者日益活跃起来时，干预对象开始包括那些长期睡在庇护所的人（之所以这样细微地区分，是因为许多睡在庇护所的人也经常跑到街上去睡，反之亦然）。这样，在项目开始出现的时候，一些人提出，那些居无定所只能在短期内住在亲戚、朋友、有时甚至陌生人家里的人也应包括在无家可归者的范畴内。对于另一些项目方而言，无家可归人口同样也包括许多住在贫民窟的个体，他们需要每天或每周付房费，而且没有租期或其他合同规定的保障。项目对象定义的这些改动导致了无家可归者项目需要服务的人群范围及其界定的变化。

对象是什么

社会项目的对象通常是个体，但也可以是群体（如家庭、工作组、组织机构）、与地理和政治有关的区域（如社区），或者物质单位（如房屋、道路系统、工厂）。不过，不管对象是什么，在需求评估开始时，都必须清晰地定义所要讨论的单位。

就个体情况而言，经常根据社会和人口特征、地点以及他们的问题、困难、个人状况等因素来辨认对象。某个教育项目的对象被指定为年龄在 10 ~ 14 岁低于在校一般年龄标准 1 ~ 3 岁的儿童。一个为母亲与婴儿提供照顾的项目的对象，是年收入低于贫困线水平 150% 的孕妇和婴儿的母亲。

当对象是集合体（如团体或组织）时，经常根据组成这些集合的个体特征来定义对象：他们非正式或正式的共同财产以及他们共同面对的问题。譬如，某个教育项目在组织层次上的对象是那些至少有 300 个学生的小学（从幼儿园到八年级），在这些学生中，有 30% 的人满足联邦免费午餐项目的要求。

直接或间接对象

对象是直接还是间接，取决于对其服务的提供是即刻的（直接的）还是最终的（间接的）。多数项目是指定直接对象。例如，在医疗项目中，很显然的情况是，患有指定病症的人能够直接得到医疗救治。然而，在有些情况下，因为经济原因，或者是出于可行性考虑，计划者会通过作用于间接人群或状况（将反过来影响预期的目标人群）的方法设计项目来间接地影响目标人群。例如在某个农业发展方案中，从社区中挑选一些有影响的农场主参加一个强化培训项目，目的是要这些农场主回到社区之后，把他们学到的知识传达给其他的农场主——项目的间接对象。

明确指定对象

初看起来，对目标人群规模和分布状态的明确指定似乎很简单。尽管关于对象的定义很容易表达出来，但是很难将这些定义应用到需求评估和项目设计

等更加细致的工作中。几乎没有什么人类社会的问题,能够根据正在经历此问题的简单而明确的个体特点,来轻易并令人信服地描述出来。

举一个例子:在一个给定的社区内,癌症患者是什么样的人?首先,对这个问题的回答取决于在计数时是否只是统计常住人口还是也包括临时人口(在任何有大量度假者的社区如奥兰多和佛罗里达,这一决定将特别的重要)。其次,康复的人是被计入?还是把五年内没有复发的人从估计中排除?第三,患有癌症者被定义为只是已经确诊的人还是包括那些症状还没有被发现的人?所有癌症患者,不论癌症种类或严重程度,都被包括进去吗?对于任一项目,虽然都应该而且可能对这类问题给出相应的答案,但是以上讨论已经说明评估者在准确描述项目目标人群上存在着诸多困难。

对象边界

对对象的详细描述需要确立区分对象的分界线,即在使用这个描述时决定谁或什么可以包括在内或者被排除在外的规则。在详述目标人群时,把定义规定得过宽或者过窄都是冒险之举。例如,把罪犯看成是违反任何法律和管理规定的任何人的描述就是无效的,只有圣人从未违反过法律或规定,不管有意或无意。这个关于罪犯的定义太概括,把轻微的或者严重的犯罪分子与罕见的违法者即习惯性的重罪犯混为一谈。

有时,定义也会太过局限或者不够综合,扼要地讲,就是几乎没有人能落在目标人群中。设想一个改造刑满释放重罪犯的项目设计者决定项目的对象只包括那些从未嗜毒或者嗜酒的人。然而在释放出来的犯人中,药物和酒精滥用现象太过普遍以至于只有很少的一部分人符合这种规定。而且,因为那些有长期拘留和定罪历史的人更可能是药物和酒精滥用者,这种定义就将把这些最需要改造的犯人排除在项目对象的范围外。

除了指定合适的分界线外,有用的对象定义还必须切实可行。如果一种特征很难被观察到或者现有的资料没有包含对这种特征的测量,那么依据这个特征做出的描述(如把一个工作培训项目的对象定义为那些对工作培训持积极态度的人)几乎不可能付诸实施。而非常复杂的、需要大量详细信息的定义,同样也很难获得应用。复杂的描述与多限定的描述相近,而且也会冒同样的风险。这样定义的生产合作社成员:既要求至少种植过两季大麦,又要求有一个年轻儿子。即使不是没有可能,也很难找到。

关于对象描述的不同观点

另一方面,不同的人会对目标人群有不同的定义,这些人涉及专业人员、政治家和其他项目方,当然也包括潜在的服务接受者。例如,不同级别政府立法者的观点之间就有矛盾和分歧。在联邦层次,国会计划通过鼓励各州投资各种赈灾措施,如提高对发生水灾的平原地区的土地使用管理和制定降低灾害风险的法规来减轻政府对自然灾害的财政负担。从联邦的角度来看,目标对象被定义

为在100年间发生洪水泛滥的所有地区。因为联邦政府必须考虑在美国的所有发生水灾的平原地区,因而他们认为这样的洪水在某一地方将每隔几天发生一次。然而,正如其名,百年洪水每个世纪里在某一地方都只发生过一次(平均而言),所以从地方的角度而言,或许根本就不把某水灾平原地区看作是合理的干预对象,因而地方政府也会强烈反对承担该类项目的负担。

在项目的设计中,项目主办方和其他利益相关方也会出现观点分歧。关注于改善穷人住房质量的项目设计者对于房屋质量的定义,与那些住在这些房屋中的居民的定义会有很大的不同。所以,在不合格住房翻新项目中,对目标人群构成的定义,甚至会包括那些认为房屋足够好的居民。

尽管需求评估不能确定有关项目对象的哪个观点是"正确"的,但能够帮助消除群体之间的冲突。要完成这一点,就要调查所有重要项目方对对象的定义,并确保这些定义在决策过程中(确定项目关注点)没有被遗漏。从不同观点中搜集的关于需求的资料,将导致对目标人群和预期目标的重新定义,或者将指出放弃某个项目的合理性(特别是当不同的观点相互矛盾,而且不同的项目方强烈坚持自己的主张时)。

描述目标人群

项目意图服务的目标人群属性对于项目干预方法的选择和项目成功实施的可能性,都会产生巨大的影响。这一部分内容将讨论一些概念,这些概念构成了恰当地描述目标对象的基础,对于项目设计和实施都有着重要意义。

高危人群,需求和需要

一个公共卫生领域的概念——**高危人群**(Population at risk),对于指明对象(特别是预防性方案中的对象)是很有帮助的。高危人群指的是对于某种假设情况有极大可能性发生或经历的群体。这样,生育控制项目的高危人群通常是指处于生育年龄的妇女。同样地,用来减轻台风或飓风破坏影响的项目将把对象定义为处于这样的风暴发生的典型地带,比如,经历过灾难风险的社区。

高危人群只能用可能性的术语来定义。育龄妇女或许是在生育控制项目的高危人群,但在这个给定的年龄段里,某一妇女可能不生孩子。在这个例子中,只根据年龄一种因素来描述高危人群不可避免地导致界定范围扩大,也就是说,定义将许多因为没有性行为而无需家庭避孕努力的妇女和没有怀孕能力的妇女也包括进了对象里。

目标人群也可以根据需求状态进行定义。**需求人群**(Population in need)通常是指目前已表现出这种状态的潜在群体。需求人群通常被相当精确地定义,通常可以根据对其实际状况的精确测量而加以识别。例如,可靠而有效的测试能够判断一个人的文化程度,这些测试也可以用来指定功能性文盲的目标人群。

对于旨在削减的项目,需求人群是指那些根据年收入来判断的,低于最低限度的家庭。尽管依据某种标准(这个标准反映了在项目或政策环境中需求的社会构成),一些个体构成了需求人群,但这并不意味着他们一定需要项目或提供的服务。在某一项目中,对服务的期望和参与的意愿将界定对特定项目服务内容的需要程度,这一概念仅与“需要”部分重合。例如,社区领袖和服务提供者为睡在街上的无家可归者十分合理地确立另一个建造通宵庇护所的“需求”,但却发现一些无家可归者并不愿使用这些设施,这说明或许只存在需求而没有需要。

根据前面给出的定义,一些用于估计问题严重程度和用作项目设计基础的需求评估实际上就是高危人群评估或需要评估。这些评估能够代替真正的需求评估之用,因为技术上无法测量需求或者因为实施只针对需求人群的项目是不可行的。例如,尽管只有进行性行为的妇女才需要避孕方面的信息,但因为很难识别和找出那些有性行为的妇女,所以多数生育控制项目的目标人群仍是那些有可能怀孕的妇女,按年龄跨度来说是 15 ~ 50 岁。同样地,虽然夜校教育项目的需求群体是所有没文化的成年人,但只有那些愿意参加或者被说服参加的人才被认为是目标人群(“需要”的概念)。虽然,关于高危人群、需求人群、需要人群的区分,对于估计问题的范畴、预期目标人群的规模以及对后续的设计、实施、评估项目都是重要的。

发生率和流行率

一个有用的区分是比较发生率和流行率两个概念的区别。**发生率**(Incidence)是指在具体时间、具体地理区域或者其他指定的区域内,具体特定问题的新生数量。**流行率**(Prevalence)是指在具体时间、具体地理区域内已经存在的数量。这两个概念来自于公共卫生领域,在这一领域内这两个概念通常区分明显。例如,在一个特定月份,流行性感冒的发生率被定义为在这个月份中报道的新病例数量;而这个月内该病的流行率则是指任何时间内的患者,不管他们第一次染病是在什么时候。在健康部门,当处理的是短期疾病,如上呼吸道感染或者其他小病时,方案设计者通常对发生率感兴趣;而当处理的是无法很快根除、需要长期的管理和治疗的问题时,包括慢性疾病如癌症和临床观察的长期疾病如严重的营养失调,设计者将更关注流行率。

发生率和流行率概念也被应用到社会问题的研究中。例如,在研究对受害者犯罪的影响时,主要的测量指标是受害发生情况,即每一时间段内,新发生的受害案件(或者是受害人)的数量。同样地,在一个旨在减少酒后驾车事故的项目中,在特定时间、地点内酒后驾车事故的发生率就是达到项目要求的最好测量指标。但对于长期问题如低教育、犯罪、贫困等,流行率通常是合适的测量指标(例如贫困问题),流行率被定义为在给定时间、社区内的贫困个体和家庭人数,不论他们何时陷入贫困。

对于其他社会问题,定义目标人群时是依据发生率指标还是依据流行率指

标往往是不清楚的。在处理失业问题上,了解流行率,即在特定时间失业总人口的数量和比例是重要的。然而,当考虑到为失业者提供财政援助时,就不清楚这个定义是指在特定时间内所有失业的人,还是指在特定时间内沦为失业的人。

比 率

除了估计问题群体的规模,了解这些问题群体的人口比例也是重要的。很多时候,很有必要把发生和流行表述成比率形式来比较不同地区或问题群体。这样,在某一特定时期,一个社区内犯罪受害人的人数(发生)可以被描述成每1 000人中的比例(例如,每 1 000 人中新生的犯罪受害人的人数为 23 个)。比率或百分比在用来识别或比较不同地区、群体的目标人群状况时特别有用。例如,在描述犯罪行为的受害者人群时,按性别和年龄分组得到估计值是重要的。尽管几乎每个年龄组都遭受不同犯罪行为的侵害,但年轻人更可能是抢劫和袭击的受害者,而年长者更是盗窃案的受害者;男性成为性虐待受害者的可能性要比女性小得多,等等。通过问题严重程度和风险状况的差异来识别干预对象,可以使得项目需求评估更好地得到完成,从而使项目实施适用于不同的人群。

在许多情况下,描述年龄比和性别比不仅是传统作法,而且非常有用。在有明显亚文化区别的社区中,种族群体归属、民族群体归属和宗教群体归属成为重要的特征分解所依据的共同因子。其他识别目标人群特征的有用变量包括:社会经济地位,地理位置和住所迁移率(见专栏 4—G 关于依据性别、年龄、种族三个变量分解犯罪行为受害比率的例子)。

描述服务需求的特征

如上所述,需求评估一个重要的作用就是提出对于给定问题和相关目标人群的严重程度和分布情况的估计。然而,由此提供的目标人群需求的具体特征也同样是重要的。通常来讲,针对特定社会问题或目标人群,某一社会项目仅仅按照统一模式送达标准化的服务仍然是不充分的。这一点之所以重要,是因为一个旨在对特定问题或需求做出回应的社会项目为了有效,还需要使其提供的服务适合问题的本土特征和需求者的独特环境;反过来,这又需要了解需求群体经历问题的方式,对于相关服务和项目的理解及权衡,以及他们在得到这些服务过程中会遇到的障碍和困难。

举例来说,一项需求评估或许会探求为什么问题会存在以及还有其他什么样的与之相关的问题。例如,关于有多少高中学生学习外语的研究能够说明在许多学校中,没有提供类似课程的服务。这样,部分问题就转化为学习外语的机会不充分。同样地,许多社会经济背景差的小学生在课堂上表现出疲倦和无精打采的现象,可以被解释为许多学生通常不吃早饭,而这又反过来反映出家庭的

经济问题。当然,不同项目方可能对问题的本质和来源有不同的看法,所以把所有这些观点都呈现出来是重要的(参见专栏4—H)。

文化因素或文化理解能形塑目标人群的基本属性,与项目到达其目标群体的有效性和提供服务的方式密切相关。例如,一个详尽的关于阿巴拉契山脉地区贫困现象的需求估计,反映出目标人群对于自足和独立问题的敏感度。慈善项目或者所提供的服务被视为施舍的项目很可能被那里贫穷但却自尊心强烈的家庭所拒绝。

服务需求中另一个重要的维度是考虑目标人群在使用服务时所遇到的困难。这些困难源于交通问题,有限的服务时间,缺少对儿童的看护以及其他许多类似的障碍。一个把服务有效地提供给需求对象的项目与一个不能有效地提供服务的项目之间的区别,主要在于项目对克服这些障碍给予的关注程度。为参加者提供儿童看护服务的工作培训项目、提供对老年人送餐服务的营养项目、提供夜间营业服务的社区医疗诊所都例证了这样的方法,即把服务送达建立在对客户需求复杂性的考虑上。

专栏4—G　暴力犯罪中的受害者比例(按性别、年龄、种族分)

(每千人)年龄在12岁及以上的受害情况:2001

受害者特点	所有暴力犯罪	强奸/性侵犯	抢劫	严重袭击	轻微袭击
性别					
男性	27.3	0.2	3.8	6.5	16.7
女性	23.0	1.9	1.7	4.2	15.1
年龄					
12-15	55.1	1.7	5.2	8.7	39.6
16-19	55.8	3.4	6.4	12.3	33.8
20-24	44.7	2.4	4.2	10.7	27.4
25-34	29.3	1.1	3.6	6.5	18.1
35-49	22.9	1.0	2.1	5.2	14.5
50-64	9.5	0.2	1.2	2.0	6.2
65 +	3.2	0.1	1.3	0.4	1.4
种族					
白人	24.5	1.0	2.6	5.1	15.7
黑人	31.2	1.1	3.6	8.1	18.3
西班牙裔	29.5	1.1	5.3	6.6	16.6
其他	18.2	1.6	2.4	2.6	11.6
所有人	25.1	1.1	2.8	5.3	15.9

数据来源于美国国家犯罪受害人调查，这项调查受司法统计署委托，由人口普查办公室实施。访问涉及所有12岁及以上的被访者(涉及大约4万户家庭，合计8万人)，在过去6个月内是否受到过一次犯罪的伤害。这些调查对象构成了可以代表美国非公职人员的一个样本。

资料来源：U. S. Department of Justice , Office of Justice Programs, Bureau of Justice Statistics, *Criminal Victimization in the United States*, 2001 *statistical Tables* (Washington, DC: U. S. Department of Justice, January 2003). Accessed through www. ojp. doj. gov/bjs.

专栏 1—11 项目各方对地方健康服务问题的不同看法

在科罗拉多州三个乡村社区中进行电话调查用以发现与癌症有关的医疗服务问题。在每个社区中，研究的参与者包括①医疗看护的提供者(医生、护士、公共医务人员)；②社区中有影响的人(教师、图书馆馆长、社区机构的领导者、商业领袖)；③患有癌症的病人及其家属。尽管在服务可获得性问题上有普遍的认同，但在相关具体问题上，每个组的看法都不尽相同。

医生和医疗看护的提供者：

- 地方的医疗机构只接受付费病人，不对其他人提供服务。
- 偏远地区缺少服务。
- 因为薪水低、工作量大和病人难应付，造成医生的数量短缺。
- 没有相应的培训和设备以提供高技术的看护。

社区有影响的人：

- 人们被列在等候服务名单中几个月。
- 缺乏足够的职业技术人员和志愿者。
- 提供服务者的专业服务知识不充足。

病人及其家属：

- 有几次找不到医生。
- 现在我们有医生，但因为他的病人没有钱再支付医疗费，所以听说他将要离开这里。
- 在胸部肿瘤X光射线透视仪器到来之前，我被列在名单中已经有三个月。

资料来源：Holly W. Halvorson, Donna K. Pike, Frank M. Reed, Maureen W. McClatchey, and Carol A. Gosselink, "Using Qualitative Methods to Evaluate Health Service Delivery in Three Rural Colorado Communities," *Evaluation & the Health Professions*, 1993, 16(4): 434-447.

描述需求的定性方法

尽管很多社会学研究都采用定量方法(以数字来代表客观信息)，但是，为获得关于某种正被考虑的特定需求的详尽而结构化的资料，采用定性(非数字化的)研究特别有用。这样的研究在复杂性上包括了从对几个人的访问或分组讨论到人类学家使用的精细的人类学民族志调查等各种方法。关于这些研究应用的例子有：关于公众信仰结构的定性资料，对教育运动的有效设计有极大的帮助；又如，公众对在吸烟快感和吸烟导致健康风险之间关系的考虑是什么？一个

好的教育项目与这些看法相适应。

定性研究(小心谨慎地实施)对于发现这类项目的过程信息特别重要。例如,关于高校学科问题的常人方法研究,不仅能够提供关于这些学科问题分布有多广的说明,而且能够指出一些学校的学科问题比其他学校少的原因。关于这些学校之间区别的研究发现,能够为项目的设计规划提供一些有价值的信息。再比如,关于家用能源消耗的定性研究揭示出这样的事实,即很少有户主知道他们设备的耗能特点;不知道如何使用能源,这些户主当然就不能很好地提供有效的节能方案。

为获得一项社会问题的丰富资料,一个普遍且有用的技术是**专题小组**(Focus group)调查方法。在一个主持人的组织下,专题小组把一些事先选定的人召集起来讨论一个特定的问题或主题(Dean,1994;Krueger,1998)。专题小组参与者通常包括有学识的社区领袖、服务机构的领导、这些机构中与客户直接接触的员工、支持团体的代表、直接经历问题和有服务需要的人以及相关的其他项目方。经过对个体的仔细挑选和组合,一个人数适当的专题小组能够提供大量的描述资料,即社会问题的本质和细微差异以及经历此问题群体的服务需求(专栏 4—I 提供了一个对于需求评估专题小组有帮助的调查书)。为获得需求评估资料的一系列专题小组技术可参见维特金和阿尔舒德的著作(Witkin and Altschuld, 1995)。

专栏 4—I 需求评估专题小组调查书

专题小组的调查书是一系列关于主题和开放式问题的清单,在专题小组会议上用来引导小组座谈。调查书应该①以逻辑顺序安排题目,从而使得一个问题建立在另一个基础上;②提出开放式的问题:这些问题具有吸引力,并且适合参与者;这些问题能够引起专题小组做出共同的回应;③划分出若干个易控制的题目组,这些题目组在限定的时间内一次讨论完一组。例如,下面是在由低收入妇女组成的专题小组(目的是探索阻碍低收入妇女获得家庭援助服务的各种障碍性因素)中使用的调查书:

- 引言——问候语;阐述会议的目标;填写姓名卡片;介绍观察者,背景规则以及专题小组如何运作(10 分钟)。
- 参加者介绍——只介绍名字;参加者住在哪里、孩子的年龄、可以获得哪种家庭援助服务以及要多久才能获得其他服务(10 分钟)。
- 介绍对阻碍服务获得的障碍性因素的看法——询问参加者关于什么是阻碍服务获得的最大障碍的看法(探索有关交通、服务机构员工的态度、管理规定,等候名单等因素);提供给她们的服务是否中断或者是否无法获得她们想要的服务(30 分钟)。
- 探索她们认为的最重要的障碍性因素背后的原因(20 分钟)。
- 询问关于在未来采取什么样的措施来克服这些障碍的看法——什么能够使得或者已经使得进入更容易,并停留在服务圈内?(30 分钟)。
- 汇报并总结摘要——主持人总结、澄清,其他的评论和询问(10 分钟)。

资料来源:Susan Berkowitz, "Using Qualitative and Mixed-Method Approaches" in *Needs Assessment: A Creative and Practical Guide for Social Scientist*, eds. R. Reviere, S. Berkowitz, C. C. Carter C. G. Ferguson (Washington, DC: Taylor & Francis, 1996), pp. 121-146

所以，在需求评估中任何使用主要知情者的方法，都必须仔细筛选那些看法会被纳入考虑的人或团体。在需求评估中，识别主要知情者的一个有用方法就是**滚雪球抽样**。这项技术要求通过合理的手段能找到最初的几个合适的知情者并对其调查。然后，请他们去找出其他他们认为对讨论的问题很熟悉的知情者。联系并询问了这些知情者后，接下来，又去找另外的知情者。当这个过程再不能找出有关的新的知情者时，就很可能把所有符合主要知情者条件的人全部识别出来了。

因为那些活跃并参与社区公众感兴趣事务的人，大多都相互认识，所以滚雪球方法对于社会问题的主要知情者调查极为有效。一个特别有用的知情者组（需求评估中不应该被忽视）是项目的现有对象和新项目的潜在对象的代表。这样的组，理所当然，对问题的特点和有关需求（这些需求是那些被问题影响的人所体验的）相当熟悉。尽管他们对问题影响程度的说明不一定特别合适，但是他们是以下方面的见证者：问题影响的程度以及问题的什么方面最亟待解决。专栏 4—J 列举了潜在服务受益者的独特观点。

因为定性与定量方法具有各自独特的优势，一种有用且经常被采用的策略就是分两步处理需求评估。第一步，探索性阶段使用定性研究方法来获得关于问题本质的详尽资料（e. g. Mitra, 1994）。第二步，根据这些资料得出估计，以设计出一个更定量化的评估方案，这个评估方案能对问题的严重程度和分布状态提供可靠的判断。

专栏 4—J 无家可归者提出他们自己对服务的需要

因为对无家可归者提供的帮助不只是提供临时庇护所，所以了解无家可归者自己对服务需求的观点是重要的。从纽约市的庇护所访问的 1 260 个无家可归者代表样本中得到的回答显示，他们有多种需求，且不是单一服务能够轻易满足的。在 20 个项目中，每项服务需求的百分比如下：

能找到住所	87.1	药品问题	18.7
有稳定的收入	71.0	学会更好地与别人相处	18.5
找工作	63.3	精神与情感问题	17.9
提高职业技能	57.0	学会怎样保护自己	17.6
学会如何获得机构提供的资源	45.4	学会怎样阅读和填写表格	17.3
获得公共援助	42.1	法律问题	15.0
健康与医疗问题	41.7	喝酒问题	13.0
学会怎样理财	40.2	在城镇周围走动	12.4
和家人相处	22.8	获得退伍军人津贴	9.6
获得 SSI/ SSD	20.8	警察的问题	5.1

资料来源：Daniel B. Herman, Elmer L. Struening, and Susan M. Barrow, "Self-reported Needs for Help Among Homeless Men and Women," *Evaluation and Program Planning*, 1994, 17(3):249-256

小　结

- 在评估研究中,需求评估试图回答关于项目所意图改善的社会状况和需求的问题,或者决定是否有实施新项目的必要。更宽泛地讲,需求评估可以用来识别、比较、优选在同一项目内或不同项目间的需求。
- 对社会问题的充分诊断和对项目目标人群的识别是设计和实施有效项目的前提。
- 必须认识到社会问题并不是纯粹的客观现象,相反,它们是社会建构。评估者是重要的辅助角色,他们能够帮助政策决策者和项目主管在具体项目中界定社会问题。
- 为了指明问题的范围和分布状况,评估者可以从现有资料,如美国人口普查中搜集和分析资料,或者利用正在运用中的社会指标来辨认趋势。因为需要的资料经常不能从上面提到的来源中获得,评估者经常需要对社会问题亲自调查。有用的方法包括:机构记录的研究、一般性的调查、人口普查和主要知情者调查。以上每一种方法都有其优缺点,例如,主要知情者调查可能相对容易实施,但其可靠性却易受到怀疑;机构记录通常可以显示出对某种服务存在需求,但却不够全面;调查和人口普查能够提供有效的、有代表性的资料,但其费用高且对技术要求也很高。
- 对未来需求的预测与需求评估有很大相关,但涉及许多复杂的、技术性的活动,通常由专家来实施。在运用预测时,评估者必须仔细检查预测所依据的前提假设。
- 关于项目对象的数量和特征的恰当定义和准确资料在整个项目过程中(从最初设计到实施的所有阶段)是至关重要的。项目对象可以是个体,也可以是群体、地理区域或物理单位,包括项目的直接或间接对象。
- 好的项目干预对象说明,能够建立起恰当的分界,使得项目能够准确地定位目标人群并且易于实施。在定义对象时,必须考虑到项目各方的不同观点,在对象定义中有用的概念有:发生率和流行率,高危人群,敏感性和特殊性,需求和需要,以及比率。
- 为了项目设计和评估的目的,获得社会问题的本土特征和项目服务需求者的独特环境的详尽资料是重要的。这些资料通常通过定性方法,如常人方法研究或专题小组(在项目各方和观察者中挑选有代表性的人组成),能很好地获得。

基本概念

专题小组(Focus group):因为对某个感兴趣话题的认识和看法而被召集起来,在一个主持人的协助下,对话题进行讨论的专门小组。这种讨论通常被记录下来,并且用来确认一些重要的议题或者建构关于这个话题的观点和经验的描述性梗概。

发生率(Incidence):在具体时期、具体区域内,具体问题的新生数量。

主要知情人(Key informants):指一些人,他们本人或职业的位置使得他们对某个社会问题或对象人群的特性和范围有较多的了解,并且在需求评估中能够表达他们的观点。

高危人群(Population at risk):具体区域内的个体或群体,拥有保持或发展某个具体状态的极大可能性。

需求人群(Population in need):在某个具体区域内,表现为某种问题状态的个体或群体。

流行率(Prevalence):在具体时期,具体区域内,某个状态已经存在的总量。参照发生率。

比率(Rate):某个具体状态的发生或存在情况被表述成在相关人口单位中的比例(例如,每1 000个成年人中的死亡人数)。

抽样调查(Sample survey):从一定群体中抽取一定量的样本,进行系统分析,其结果可以通过统计推断来反映整个群体的特征。

滚雪球抽样(Snowball sampling):一种非等概率的抽样方法,通过被访人向研究者不断推荐新的研究对象来完成研究,这个过程可以一直持续到没有新的对象可以被挖掘。

社会指标(Social indicator):用来追踪一段时期内社会状态的定期测量工具。

5 项目理论的表达与评估

在第 3 章,我们主张评估者分析项目理论,并将其作为确定潜在重要评估问题的一种辅助工具。在本章,我们提出项目理论议题,但并不是将其作为界定评估问题的框架,而是将其作为项目本身的组成部分来评估。

项目所要解决的社会问题通常是十分困难和复杂的,即使付出很大的努力,也很有可能只带来很小的改进。项目理论是一种概念体系,即关于为实现预期的社会效益应该做些什么的设想,是项目执行的基础。

好的项目理论使我们知道要怎样做才能达到项目目标,哪些事情是项目必须做的。相反,较差的项目理论就算执行没有问题,也不能产生预期的结果。项目评估的一个重要方面,就是评价项目理论的优劣,评价项目理论是否有清晰的逻辑,是否提供了合情合理的改善社会状况的方案。待评估的项目理论必须首先得到清晰和完整的表达,这样我们才能开展评估工作。本章介绍了如何表达项目理论以及评价这些理论优劣的方法。

前纽约州州长库莫(Cuomo)曾经引述其母关于成功的看法,即清楚自己到底要干什么,并且扎扎实实地去做这些事。成功的社会项目也应该具有这样的特征。如果给定明确的需求,项目决策者必须制订一套能够满足这些需求的方案,并且执行这套方案。在本章,我们将介绍一些评价项目构思质量的概念和方法,即项目理论。在接下来的一章里,我们将讨论如何评价项目执行的质量。

项目理论无论以详细的项目计划和理由来表述还是仅隐含在项目的结构和活动中,都解释了项目为什么采取那些行动,并且给我们提供了合理性的基础:只要按照项目的规定行动,就可以得到所需要的结果。在核查项目理论时,评估者往往发现理论不太令人信服。糟糕的项目设计的缺点体现在其所提供的达到预期目标的隐含假设中。发生这种情况主要是因为在设计项目时,对于提供细致明确的项目目标和执行步骤的重要性认识不足。有时可能是因为项目所处的政治背景不允许更进一步的计划,但是很多情况是,常规的项目设计并没有深入考虑到其隐含项目理论的性质和可行性。人们往往是凭着他们已经习惯的行为方式做事,各行有各行的规矩,因此,项目设计首先是如何实际应用的问题,而不单是服务和问题之间简单的逻辑分析。

例如,在针对诸如酗酒、吸毒、青少年性行为、未婚母亲和犯罪这样的越轨人群的防治项目里,人们往往认为教化和相互的忠告有助于问题的解决。尽管没有在项目中明确阐述,但还是基于这样的假设——即如果提供足够的信息和人际帮助,人们会减少这些越轨行为。这种理论对于一般人或许有效,但是经验和研究表明,即使越轨者知道如何去改正他们的恶习,也知道亲人们是如何关怀和鼓励他们,他们往往还是拒绝做出改变。因此,教化和忠告能够解决问题的假设就不是合理的项目设计基础。

有经验的评估者应该把项目所依托的理论假设作为项目的重要方面而给予相当的重视。如果项目的目标和项目所要改进的社会环境之间并不是以合理的方式联系在一起的,或者项目功能中体现的预期和假设不能够提供获得改进的可行途径,那么项目本身也不大可能是高效的。

评估项目理论的第一步是阐述项目理论,即建构一套明确的概念、假设和预期作为构造和执行项目的逻辑基础。项目很少向评估者提供充足的、有助于评估的、直接明确的项目理论的阐释。项目理论往往隐含在项目的结构和执行过程中,在相关的项目文件里很少能找到完整的项目理论。即使能从项目的书面材料中找到项目理论的陈述,也往往是出于申请资金或公众支持的考虑,与实际的项目进程有很大的出入。

因此,项目理论的评估首先要求评估者通过各种途径对项目理论进行综合和阐释,使之适合于评估的要求。据此,本章将围绕两个主题来展开:①讨论项目评估者如何阐释既符合项目各方实际理解又有利于展开评估的项目理论;②然后讨论如何评价这些明细化的项目理论的质量。我们从对一种视角的简要描述开始讨论,这个视角提供了评估项目理论的最为完善的方法。

可评估性评价

最早关于阐述和评价项目理论的系统论述是由城市大学的评估研究小组在20世纪70年代根据评估实践提出的(Wholey, 1979)。他们发现有些公共项目评估是难以甚至是无法理解的,于是便对理解评估的障碍进行了研究。通过研究,他们认为,高质量评估工作的第一步是定性评价实施评估的前提条件是否满足。沃雷(Wholey)及其合作者把这一步称为"可评估性评价"(参见专栏5—A)。

专栏5—A　可评估性评价的必要性

如果评估者和评估结果的利用者不能就项目的目标、信息的优先性和项目产出的应用达成共识,项目评估的设计就会被引到与项目政策和项目管理决策无关的问题上。如果项目本身所能提供的资源不能满足主要项目活动的需要,那么项目的目标就是不切实际的,项目不可能很好地完成。如果项目管理者缺乏关于如何达到项目目标的知识,项目也不可能达到预期的效果。因此,在正式评估活动进行之前,项目管理者改进项目资源、项目活动和项目目标是很有帮助的。如果相关的数据无法得到或者不能确定合理的成本,就无法确定接下来的评估的质量。如果政策制定者和项目管理者不能或不愿应用评估结果对项目进行改进,那么再好的项目评估也会陷入没有应用者的尴尬境地,如果这些问题得不到解决,项目评估就不会给项目进程带来任何有意义的改进。

决定公共项目或私人项目性质的这些问题,将通过高质量的可评估性评价予以解决,可评估性评价必须遵循的标准是:

- 项目目的、目标、重要的附带性影响以及信息需求的优先性必须明确地界定。
- 项目目标必须是合理的。
- 可以获得有关项目执行进程的资料和数据。
- 评估结果的应用者对评估结果如何应用达成共识。

通过可评估性评价,评估者可以明确项目的设计和项目的实际执行情况,如果必要的话,还可以通过上述的四个标准对项目进行重新设计。可评估性评价并不仅仅告诉我们项目是否可以进行有效的评估,而且可以告诉我们通过评估是否能改进项目的进程。

资料来源:Joseph S. Wholey,"Assessing the Feasibility and Likely Usefulness of Evaluation" in *Handbook of Practical Program Evaluation*, eds J. S Wholey, H. P. Hatry, and K. E. Newcomer (San Francisco: Jossey-Bass, 1994), p. 16.

可评估性评价通常包括三个基本的步骤:①对项目模型特别是项目目标进行描述。②评价项目模型界定的准确性和可测量性。③确定项目各方在项目评估中的利益所在和他们对评估发现的应用方式。项目评估者需要像民族志学者那样开展可评估性评价工作。他们必须通过访谈和观察揭示项目执行者和主要项目方对项目的实际理解。评估者常常是首先通过项目计划书和官方信息所提供的细节来了解项目,但是之后他们就要通过那些实际接触项目的人了解项目。

这样做的目的是得出一个对项目实际情况的基本描述,同时,兼顾大多数参与者关于项目的基本理解。在早期,这一步主要依靠评估者自己的判断,现在已经有很多专业人士规定了一些程序,以便其他评估者顺利地执行可评估性评价(Rutman, 1980; Smith, 1989; Wholey, 1994)。

可评估性评价往往使项目管理者和主办方都意识到改进项目的需要。可评估性评价可以发现项目执行系统的错误,譬如项目的目标群体不明确,或者是项目的活动需要重新定义。也有可能项目各方之间共同的目标很少,项目目标缺乏合适的评判标准。在这些情况下,项目评估者发现了项目设计中的问题,如果要进行有效的评估,项目管理者必须首先解决这些问题。

可评估性评价的目的在于创造一种适合于评估工作的环境,使人们就项目本身的性质及其目标达成共识以便简化评估设计。这样,评估者就会在设计评估和组织评估问题的时候受益匪浅(参见第2、3章)。专栏5—B提供了进行可评估性评价的实例。

可评估性评价要求项目各方阐述项目的设计和项目的逻辑(即项目的模型),而这也是构造和评估项目理论的要求(Wholey, 1987)。实际上,可评估性评价方法提供了描述和评价项目理论的一系列概念及程序,通过深入研究,可以知道项目设想做什么以及为什么要这样做。因此,我们接下来会就界定和评价项目理论问题进行细致的讨论,而把重点放在有关可评估性评价的实践方面。

专栏5—B 阿巴拉契亚地区委员会的可评估性评价

城市大学的评估者曾经在一项健康和儿童发展项目中与阿巴拉契亚委员会的管理者和政策制定者合作,在可评估性评价过程中,他们进行了如下工作:

- 研究了该委员会下属13个州的健康和儿童发展项目的现有资料及数据。
- 访问了5个州,并把其中的2个作为评估设计和完成评估的数据点。
- 研究了国会、阿巴拉契亚委员会(ARC)、各州、项目实际参与者和项目对象的相关文件(包括正式的行政立法、国会听证资料和委员会报告、各州的计划文件、项目申请报告、ARC的合同报告、地方性计划文件、项目资料和研究方案)。
- 走访了大约75名相关人士,他们有的在国会或委员会的核心部门工作,有的在各州的ARC委员会或健康和儿童发展项目中工作,有的在地方的规划单位任职,还有的在地方的项目执行单位工作。
- 直接到实地了解情况,走访了大约60人,包括ARC雇员、参加项目的对象以及有关的评论者。

对这些资料进行分析和综合,可以在项目活动和项目目标之间建立一套逻辑模型以说明其间的因果联系。这样就有了分析项目目标的尺度和适当性标准,新的项目设计就能够比较有效地验证项目执行的效果。这里,项目理论涵盖了整个ARC的项目模型以及一系列具体项目的模型,每个模型都有各自确定的项目目标。

在对项目报告进行研究的过程中,评估者要求ARC职员对备选的行动步骤做出明确的说明,研究过程包括城市大学的评估者与ARC人员就项目目标和项目模型进行一系列讨论。在每个阶段,

评估者和ARC人员们都期望就项目理论的妥当性、各个项目对象的重要程度和评估所涉及各个问题的解答程度达成一致的意见。

ARC最后改进了项目的设计,决定系统地督导所有健康和儿童发展项目的执行情况并且评价那些经过改进的项目的效率。在13个州中有12个做出了相应的改进,现在,项目计划书比以前更清晰地阐释了项目的设计,他们认为项目已经得到了很大的改进。

资料来源:Joseph S. Wholey, "Using Evaluation to Improve Program Performance," in *Evaluation Research and Practice: Comparative and International Perspectives*, eds. R. A. Levine, M. A. Solomon, G.-M. Hellstern, and H. Wollmann (Beverly Hills, CA: Sage, 1981), pp. 92-106.

描述项目理论

长期以来,评估者已经意识到将项目理论作为工作基础的重要性,这个基础将帮助阐述评估问题、设计评估研究以及解释评估结果(Bickman, 1987; Chen and Rossi, 1980; Weiss, 1972; Wholey, 1979)。然而,在具体使用的过程中,评估理论却有过许多不同的名字,例如逻辑模式、项目模式、产出草图、因果图、行为理论,等等。在更多的情况下,对于项目理论如何被描述到最佳程度,并没有统一的意见,因此,我们将描述一个在评估活动中被发现比较有用的框架。

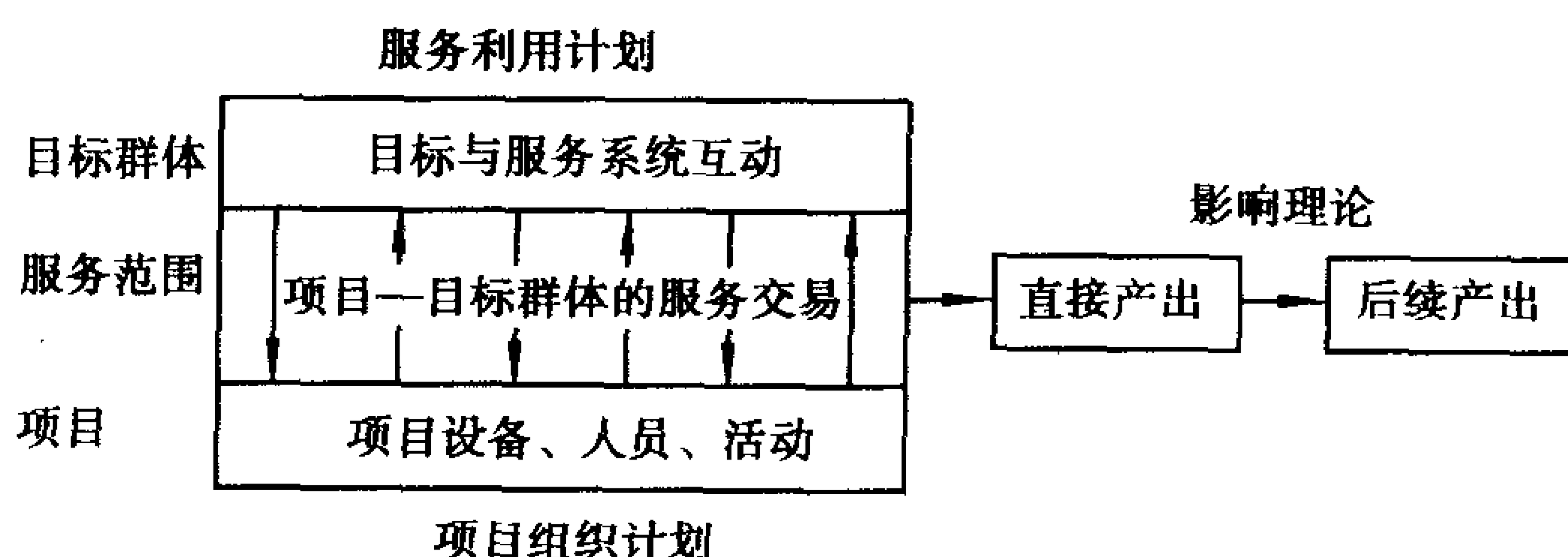
为了达到这个目的,我们将介绍一个典型的社会项目作为分析对象,描述项目操作和项目服务目标人群之间的关联(参见专栏5—C)。这些关联性事务或许包括健康机构为有不良饮食习惯的妇女所做的咨询,社区中心为处于危险年龄的年轻人举办娱乐性的活动、为当地的居民提供的教育性讲演,在诊所内张贴的营养知识海报,散发的有关加强政权和税法的传单,为老年人提供的上门送餐服务,或者是任何服务性的合同。对于这种有目标的项目业务,一方面我们将项目作为有组织的实体,其中包括各种辅助性工具、人员、资源、活动,等等。另一方面,我们还关注目标参与人群,他们处在不同的环境中,有着不同的行为表现,包括他们的环境和经验,这些都与服务送达系统有着密切的联系。

这个简单的框架指明了三个不同的、但互相联系的理论组成部分:项目影响理论、服务利用计划及项目组织计划。**影响理论**(Impact theory)就是有关项目行动导致变迁,并在变迁中改善社会状况的项目假设。其中,最重要的是项目—目标人群之间的关联,其中包括项目达到预期效果所需使用的运作方式。这样的影响理论有时候很简单,譬如向人们说明滥用麻醉品会引起副作用,就会使中学生对这些信息有更多的了解;有时也会很复杂,譬如让八年级的学生对自然现象有更深刻的理解;有时候又不那么正式,譬如告诉人们为老年人提供上门送餐服务会改善他们的营养结构;有时候却很正式,譬如用经典环境理论来对付恐惧症。然而无论其本质如何,某种类型的影响理论组成了社会项目的核心。如果理论中有关如何达到预期变化的假设是由项目行为引起的,而这些假设有漏洞;或者如果这些假设很有效,但并没有很好地被组织起来,那么所期望达到的社会

收益便实现不了。

为了获得项目影响理论所期望的变化，项目首先必须为目标人群提供计划好的服务。如果我们从目标人群的角度来衡量项目，那么注意力就应该集中到提供服务上，看是否准确地为目标人群提供了服务，以及整个服务送达的程度。关于这些问题的假设和期望是项目理论的重要组成部分，我们将其称为项目**服务利用计划**(Service utilization plan)。当服务已经完成或目标人群不再需要服务时，就意味着这种关系的终结。例如，对于试图提高人们对艾滋病的了解的项目服务利用计划而言，项目活动或许只包括在地铁内张贴告示以让人们阅读。但是对于具有多种活动的预防艾滋病项目而言，其基础或许是看没有得到服务人员帮助的吸毒高危人群是否在街边诊所接受了检测和获得了相关的信息，如果是，这些吸毒者便能得到项目试图提供的服务。

专栏5—6 项目理论概览



当然，项目必须用特定的方式组织起来，使其能够真正地为人们提供服务，并同时产生人们所需的收益。因此，项目理论的第三个重要组成部分与项目资源、人力、执行情况和总体组织有关，这就是**组织计划**(Organizational plan)。组织计划的原理是：如果项目有这样或是那样的资源、设备、人员等，并且能够被合理地组织和利用，产生相应的活动和功能，那么，可行的组织便会形成，并能拥有发展和/或保持服务供给系统和服务利用的能力。项目组织理论的要素包括类似于这样的假设：项目经理应该在社会工作方面取得了硕士学位并至少有五年的工作经验；至少雇佣20位项目经理，某部门应该有一个智囊团为当地的商界人士提供信息，在每个地方都应该委派一个执行人，应该经常与公共健康部门保持联系，等等。

按这样的主旨，充足的资源和有效率的组织是发展和保持服务系统的重要组成部分，这样的系统使服务利用成为可能，从而使目标人群能获得这些服务。项目组织及其所支持的服务送达系统是直接接受项目执行人员领导的项目组成部分。这两个方面联合起来，通常被称为项目过程，同时，项目过程所依赖的假设与期望则被称为项目**过程理论**(Process theory)。

有了这个综述之后，我们将对项目理论构成的细节进行讨论，尤其是看评估者如何才能获得可行的项目理论，以及如何用项目理论来分析项目、发现潜在的重要评估问题。

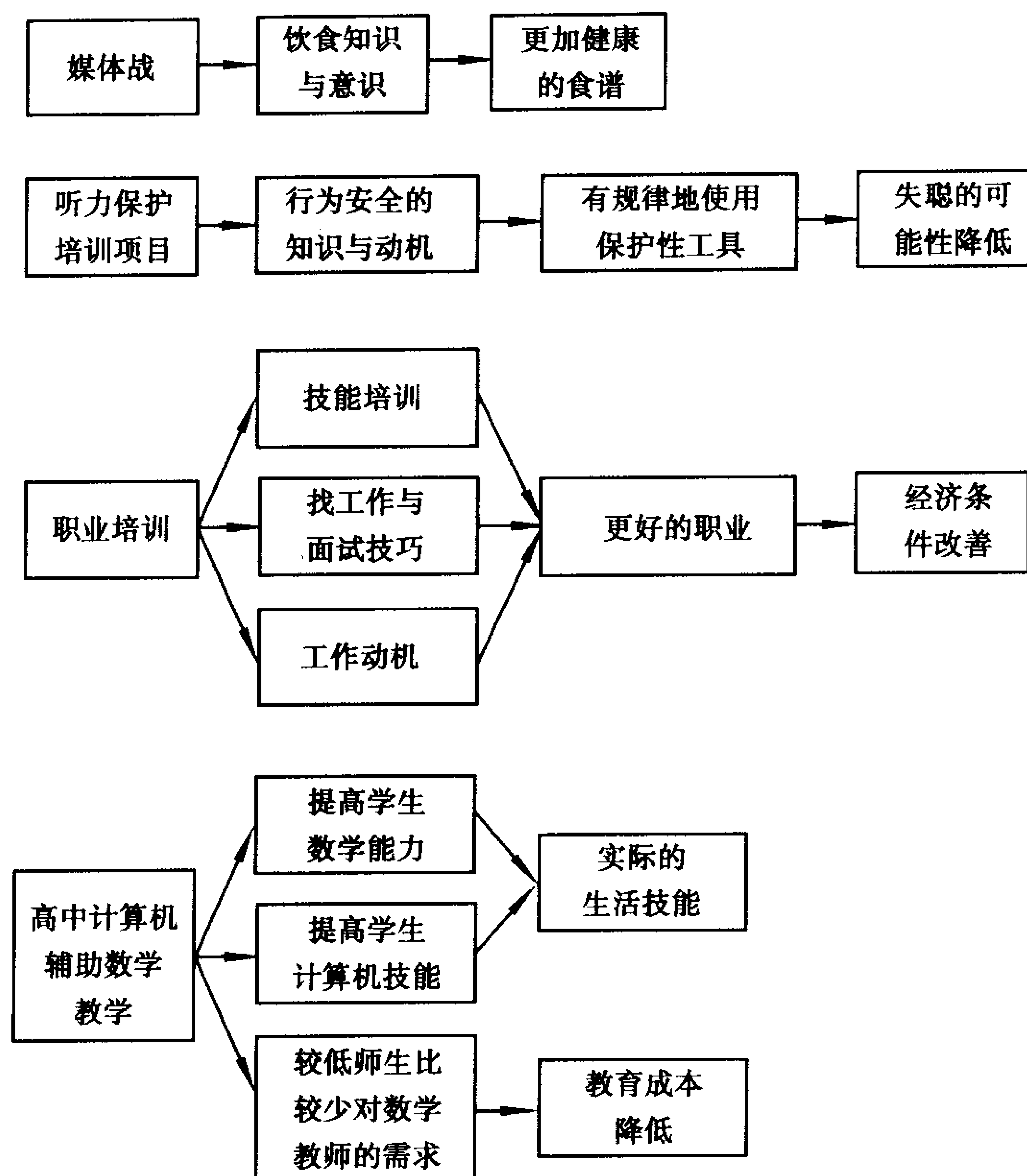
项目影响理论

项目理论是因果关系式的,它被用来描述特定项目活动(诱因)和特定社会收益(效果)之间的因果关系。因此,评估者在展示项目影响理论时,总是使用因果关系图,显示原因和结果之间的联系,并将项目活动和期望结果之间的联系推测出来(Chen, 1990; Lipsey, 1993; Martin and Kettner, 1996)。因为项目很少能被彻底地执行,因此不可能对社会环境进行直接控制从而达到改善环境的目的。在通常情况下,项目需要尝试改变一些重要的、但可更改的环境的某些方面,进而导致更多社会环境的改善。

因此,最简单的项目影响理论常常被概括为基本的"两个步骤",首先,服务改变了一些中间环节的状态,比方说动机,然后再帮助改善人们所关心的社会状况(Lipsey and Pollord,1989)。例如,某项目或许无法让人们放弃酗酒,但是可以尝试改变人们对酒的看法以及喝酒的动机,进而帮助人们避免酗酒。复杂的项目理论在项目执行和获得社会收益两个过程之间会有更多的步骤,或者项目所执行的程序不止一个。

任何项目影响理论都会包括一些这样的因素,或者是原因或者是效果,在这些因素之间形成的因果联系反应了事件的发生链,这条链连接着项目最初的活动和最终改善了的社会条件(参见专栏 5—D)。按照项目影响理论图所示的因果流程图来看,在起始性项目活动之后的每个事件都是一个结果。紧接着起始性项目活动之后发生的是最直接的结果,这常常被称为最接近结果或是直接结果,如专栏 5—D 中饮食知识与意识;而那些在接下来出现的则被称为后续的或是最终结果,如专栏 5—D 中更加健康的食谱。在成功地出现了最直接的结果之后,项目影响理论为那些在迟些时候将会出现,而且在通常情况下更为重要的结果提供了理论支持。

专栏 5—D　项目影响理论图解

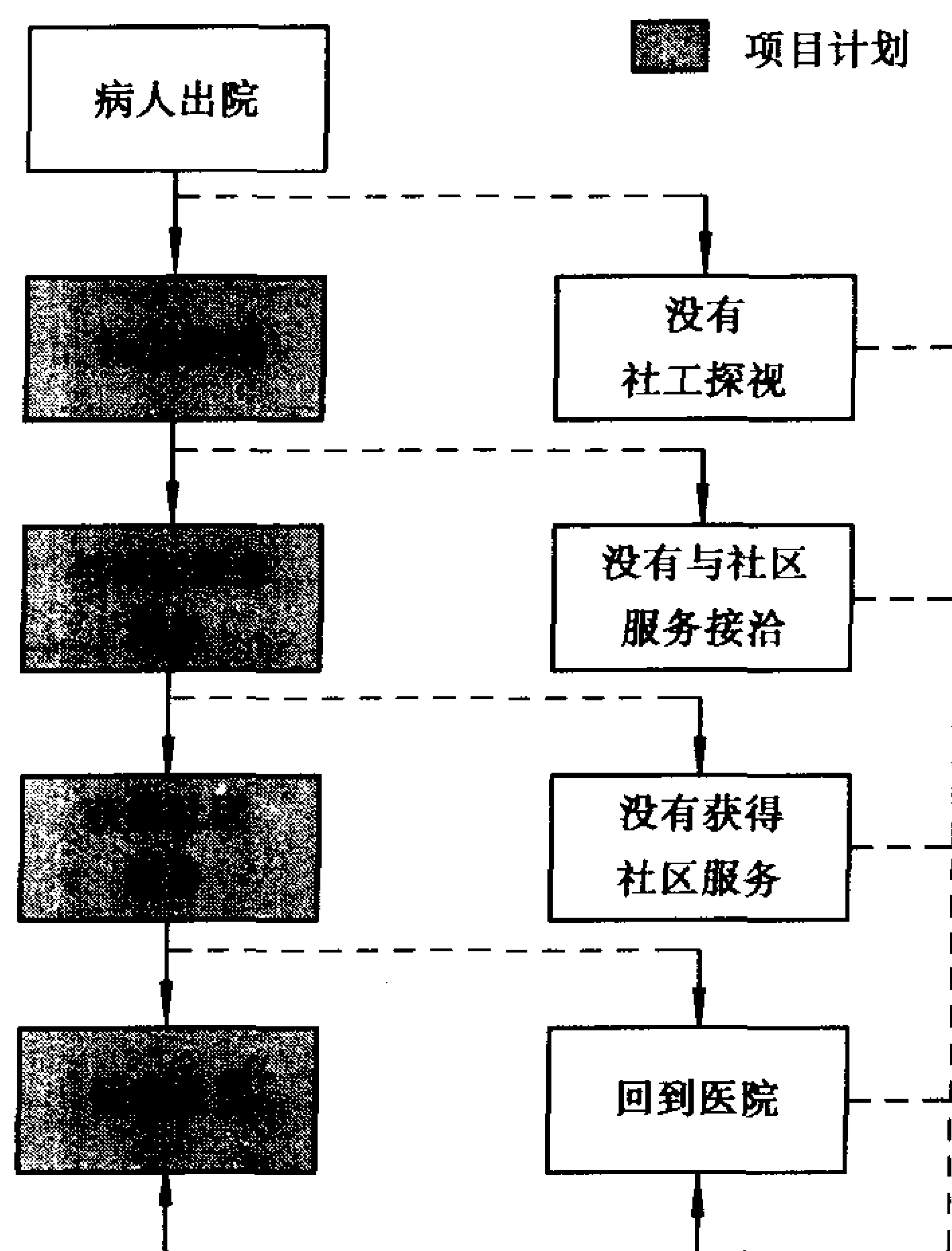


服务利用计划

清晰甚至不太正式的服务利用计划将注意力集中在重要假设推断方面，即服务对象怎样以及为什么加入项目、如何充分得到项目影响理论所预期的服务。服务利用计划从目标人群视角和他们在项目中所经历的事件出发，描述项目和干预对象之间的关联（交易）。

项目的服务利用计划可以用流程图详细表述，从而跟踪项目目标的各种路径，而这些路径的起始点又可以是本项目之前的某个点。专栏 5—E 介绍了一个服务利用计划的简单例子，其内容是给精神病人提供服务。这种流程图要达到的目的之一，就是避免遗漏项目目标。例如，在专栏 5—E 中显示的出院后社区护理服务项目中，我们可以发现，在目标人群中，没有在医院接受过正式精神治疗的病人就不会获得项目为他们安排的、来自于社会工作者或指定机构的探视，或者根本就不接受任何服务。

专栏5—E 照料出院病人服务项目流程图

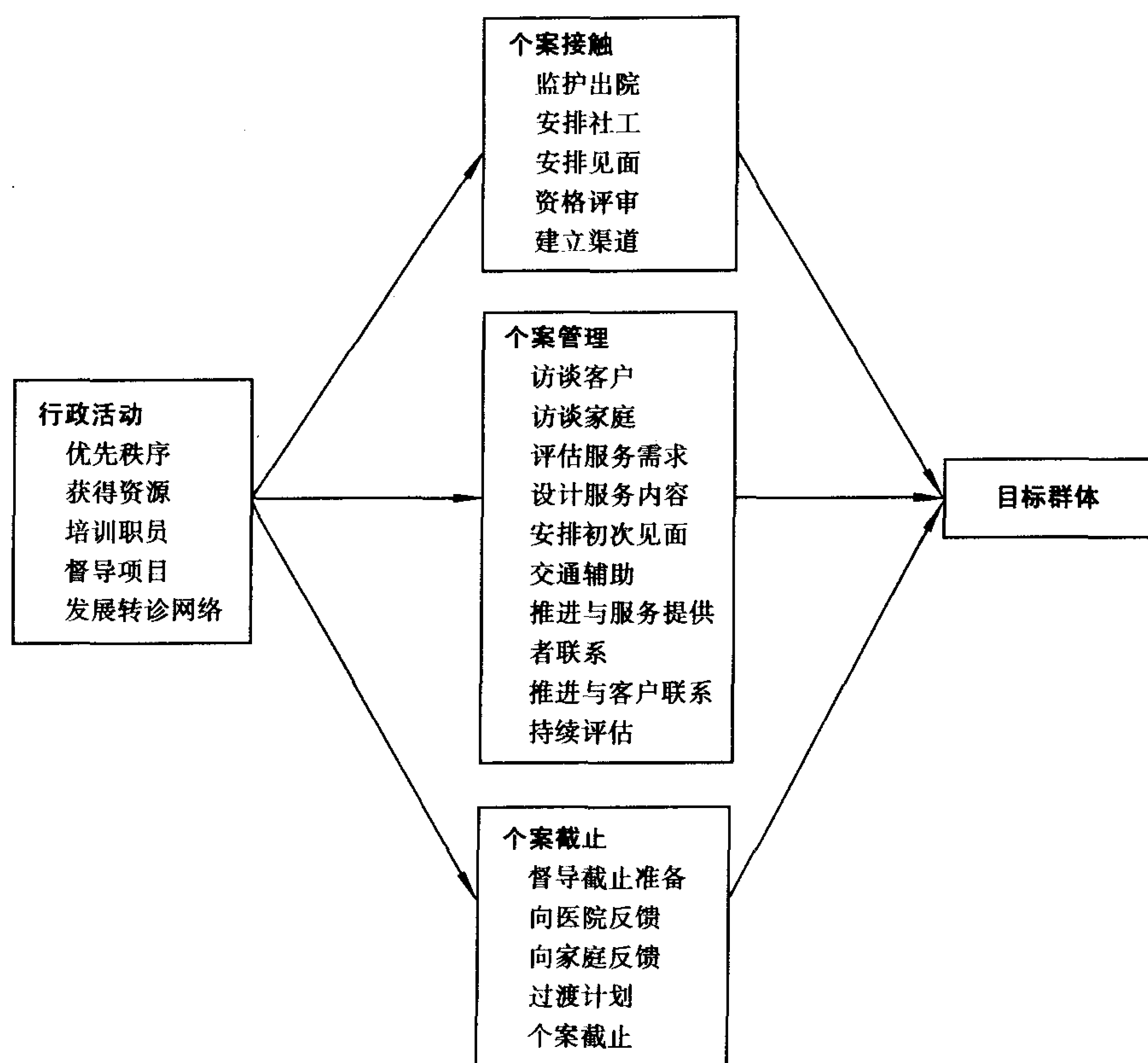


项目组织计划

组织计划是从项目管理中总结出来的,内容包括项目想要发挥的功能、组织活动以及实现功能所需的人力、财力和物力资源。核心内容是项目服务,在项目目标—项目服务转换中,这些具体服务使项目能发挥其应有的角色作用,从而创造出社会收益。不过,组织计划也必须包括那些能为组织提供重要先决条件和支持的功能,从而使组织具有提供主要服务的能力,例如,资金积累、人员管理,设施获取和维护、政治交涉,等等。

评估者可以通过许多方法来描述项目的组织计划。如果我们将注意力集中在项目目标—项目服务转换上,组织计划的第一个成分就是描述项目目标及其针对目标所提供的具体服务:服务的内容是什么,应该提供多少服务,为谁提供服务,按照怎样的程序提供服务。组织计划的下一个内容是,描述这些服务活动所需的资源和必要的功能性保障。例如,需要由有资格凭证和技能的充足人力资源来提供后勤服务、适当的工具和设备、资金、监管、办事人员的支持,等等。

专栏 5—F 照料出院病人的组织示意图



至于项目理论的其他部分，在通常情况下，用图表或流程图来描述项目组织计划是很有用的。专栏 5—F 提供了一个例子，该图表显示的是为精神病人安排出院后服务项目的重要组织成分，这个例子的服务利用宗旨已经在专栏 5—E 中描述得很清楚了。呈现项目组织计划的最常见方法，用术语表述就是投入（反应了项目可利用的资源和限制条件）和活动（项目预期提供的服务）。在完整的项目逻辑模型中，接受服务（服务利用），即项目产出，与预期的项目产出相关。专栏 5—G 展示了一个从一本广泛使用的业务手册摘取的典型逻辑模型，这本业务手册由美国联合慈善总会（The United Way of America）提供。

专栏5—6　对少年母亲进行为人父母教育的逻辑结构

投　入	活　动	产　出	产　出			
			最初的	中间的		最终的
项目机构提供MSW项目经理、兼职的RN(Reqistered Ner-se，注册护士)指导员、国家认可的教育手册、录像带和其他的教育工具	从怀孕三个月之后到一年期满(子女出生)，每周都会为怀孕少女安排两次课程(每次为期1小时)，课程内容涉及整个生育过程中需要注意的事项：孕妇及婴儿营养、安全、智力开发、照料，上课地点在高中	怀孕少女加入到项目中来	怀孕少女了解了有关生育前营养补充和健康指南的知识	怀孕少女按照指南进行适当的营养补充和健康照料	怀孕少女产下了健康的婴儿	在12个月内，婴儿接受了合适的照料，在身体、运动神经、语言能力和社会互动发展方面都取得了显著进步
项目机构和高中确定怀孕的少女，让她们加入到项目中来			怀孕少女了解了对婴儿进行适当照料、养育和进行社会互动的知识	对婴儿进行适当的照料、喂养和社会互动		

资料来源：United Way of America Task Force on Impact, *Measuring Program Outcomes: A Practical Approach.* Alexandria, VA: Author, 1996, p. 42. Used by permission, United Way of America.

构造项目理论

在一些特殊情况下,在项目的相关文件中可以找到工作人员和项目各方都比较满意的明确的相关理论。这时,我们认为项目有一个**明细化的项目理论**(Articulated program theory)(Weiss, 1997)。这种情况往往发生在那些以某种理论为指导而进行设计和制订方案的项目中。例如,在设计和执行防止青少年滥用毒品的项目中,同伴群体在戒毒行为中扮演的角色可能就是直接参照了社会学习理论及其在青少年行为中同化影响的应用。

在很多情况下,项目活动和项目提供的服务被认为是达到项目目标的合理途径,但是它们背后隐含的关于怎样达到项目的假设和解释并不是一目了然的。这时,我们就认为项目理论是含蓄的,或者用维思(Weiss,1997)的话说,是一种**隐含的项目理论**(Implicit program theory)。这种情况可能发生在为帮助那些出现感情问题的夫妇而设计的沟通项目中。尽管可以认为与专业人员讨论婚姻问题是有帮助的,但是项目中并没有明确指出通过这样的活动就能改善夫妻关系,而且并不是每个参与活动的人都认为那是有效的。

当项目理论比较含蓄时,评估者必须在分析和评估之前通过一定的方法提炼并描述项目理论。项目理论的目标是描述项目预期,即在项目决策者的设想中,项目应该做些什么,执行项目后应该带来哪些后果。经过这样的背景分析,我们再来讨论提炼和阐述项目理论的概念与程序,这是项目评估的前提。

界定项目所涵盖的范围

阐述项目理论的第一步是界定项目所涵盖的范围(Smith, 1998)。人类服务机构可以有很多项目,也可能提供很多服务。对项目唯一正确的界定往往不存在,在很大程度上,评估者构造项目理论时所要遵循的框架就是评估主办方所关注的领域。

我们可以根据决策者对于评估结果应用的看法以及他们可能会做出决策的性质来限定项目理论本身的范围。构造项目理论至少应该说明决策过程中相关决策者的权力、组织机构和决策活动。如果评估主办方只是一个地方社区的精神病治疗机构,那么项目决策和项目的边界首先应该是该机构治疗的病人和一系列的服务项目,例如对厌食症的院外治疗计划。而当评估主办方是州的精神病治疗管理机构,那么项目的范围就可以被界定为全州范围内院外治疗的效率问题,也就是州内所有的地方社区精神病治疗机构的院外治疗计划。

由于项目理论主要解释的是项目手段和结果之间的关系,所以,确定项目边界的关键是看其是否涵盖了重要的项目活动、重要事件和项目资源与项目计划中所预期的主要结果之间的联系。这就包括了某种形式的追溯,起点是一系列明确界定好的项目对象和预期社会收益之间的联系。这样,为了达成预期目标

而设计的相关活动和资源调度就成为项目的组成部分。从这个角度看,不论在地方层次还是在州层次上,厌食症的治疗计划都会被界定为各个精神病治疗机构组织的一系列项目活动。在各自权限范围内,这些机构在试图减轻厌食症的努力中扮演着重要的角色。

但需要注意的是,尽管这些为阐述项目理论而对项目进行的界定在概念上是直接的,但在实践中可能存在很多问题。项目资源、项目活动和项目目标都是复杂的,项目本身也十分复杂。而这些在理论中作为界定项目的关键,也很难被确定。所以,考虑到这种情况和评估中还要遇到的其他问题,评估者必须和项目评估主办方和主要项目方之间就项目的界定达成共识,并且使项目的界定符合评估的进程。

细化项目理论

对于处于早期规划中的项目,项目理论可以从以前的实践和研究中产生。在这个阶段,评估者的工作可能有助于建立可信的、表述完好的理论。对于已经实施中的项目来说,我们的任务是描述一个真正体现项目结构和执行的项目理论。为此,评估者必须通过与项目各方的互动找到那些在他们的行动和预设中反映出来的隐含的项目理论。大致的程序就是不断地趋近,逐步地细化。项目理论草案的产生往往要通过评估者和有经验的项目方代表不断地讨论和反馈。在各方讨论的基础上修订草案,然后再展示给项目各方,进一步征求意见,可以依次循环多次。构造项目理论的这种方法也可以用来构造项目影响理论和过程理论以及其他重要的项目要素,专栏5—H展示了评估者是如何构造项目理论的一个案例。

专栏5—H　针对适应性工作服务建构项目过程理论

田纳西纳斯维尔的罗歇尔中心发起了一项适应性工作服务项目,计划向早期痴呆患者提供强度较低的有酬工作和相应的社会交往机会。项目是基于这样的假设:即在安全的条件下进行工作不仅有利于患者的感情和智力提升,而且还可以减轻家人照顾他们的负担。评估者给出了建构该项目过程理论的程序。

建构该项目执行模型的方式是运用便笺和墙报对项目进行整版描述,第一步只有评估者和项目的最高管理者参与,要提出的问题是"当一个患者来到中心询问情况时,将会发生什么事情",评估者应该把获得的回答记录在便笺上,并且置于墙报上,接下来的工作就是确认,这些也被记录在墙报上,直到所有的已知的活动都被确认并且记录在墙报之上。当管理者确认了他所知道的所有项目活动后,评估者就要把这些便笺搜集起来。对搜集到的便笺进行讨论,找出潜在的构成,因为这些工作只是在两个人之间进行,所以可以把结论暂时搁置两周,以便行政管理者和其他管理人员对它进行充分的讨论和反馈。他们会指出评估者提供内容的缺陷和失误。在听取了从职员到行政管理者的广泛意见和讨论之后,才把结论提交到咨询委员会。委员会的成员将对结论做进一步的讨论并提出修改的意见。经过几次股东例会之后,行政管理者可能会要求重新审视项目计划的模型。这样,项目活动就明确了,同时组织对项目本身的理解也加深了。

资料来源：Quoted, with permission, from Doris C. Quinn, "Formative Evaluation of Adapted Work Services for Alzheimer's Disease Victims: A Framework for Practical Evaluation in Health Care" (doctoral diss., Vanderbilt University, 1996), pp. 46-47.

区分和发展项目理论的最基本资料来源是：①研究项目文件；②走访项目的执行者、主要项目方和相关的知情人；③选点调查，观察各种项目环境和项目功能；④社会科学文献。评估者能从这些资料中提取的以下三类信息将是很有用的。

项目目标

与项目资源密切相关的项目目标是项目最重要的特征，也是项目理论特别是项目影响理论不可或缺的组成部分。项目理论必须明确界定项目目标。但是项目文件中所界定的目标或者主要项目方所设想的目标往往与项目理论的目标不同。如果要进行有意义的评估，项目目标必须是通过一系列项目活动所能带来的产出，也就是说，项目活动与所设想的产出之间必须具有某种逻辑联系。史密斯（Smith, 1989）建议评估者一步步地提问而不是直接询问项目目标。例如，在调查主要项目活动时，评估者可能会问到这样一些问题："为什么这样做？""估计会产生哪些后果？""你怎么确定那些后果到底发生了没有？"这样，讨论就会逐渐变得集中和具体，而如果直接提问项目目标是什么，就会使讨论变得抽象，往往成为泛泛而谈。

如果对项目目标有具体陈述，就必须将其有效地整合到项目理论中去。项目的目的和目标就是与项目影响理论相关联的、项目试图改变的社会状况。譬如，在一个为缓解失业的项目影响理论中，项目目标就是项目远期产出。反之，与项目目标相关的项目活动和服务送达有助于揭示项目过程理论。如果项目目标是为工作的父母提供孩子放学后的看护，那么服务计划也就部分显现出来了。同样，如果项目目标是每周上四次阅读课，那么组织计划的一个重要部分就被确定了。

项目的功能、要素和项目活动

要恰当地表述项目过程理论，项目评估者就必须仔细而明确地确定每个项目要素、项目的功能以及特定行动与这些功能之间的关联机制。项目功能包括诸如"完全内化"、"预测对象需要"、"安排个案管理"、"吸收指定的机构"及"培训实地工作者"之类的操作。评估者一般能通过确定项目活动和对各种项目人事的工作描述来辨别各种功能。当汇集到具体项目主题之下时，这些功能就构成了项目过程理论的要素。

项目功能、项目活动和要素的逻辑关联

项目理论的一个重要方面就是不同步骤和功能是如何联系在一起。有时，这些联系只包括项目关键活动及其影响的联系。例如，在项目提供刑满释放后

的服务之前,监狱看守必须告知项目执行者有罪犯被释放了。在其他情况下,这些联系是项目活动必须协调的事情,照顾孩子或接送问题必须在职业培训计划内或诸如培训保育员这样相关的支持体系内加以安排。其他联系则需要逻辑或概念上的关联,这种情况在项目影响理论中尤为常见。因此,关于如何照顾孩子的知识与父母照顾孩子的方式之间的联系隐含着一个心理过程的假设:信息会影响人的行为。

因为这些关系往往是可见的,评估者一般用图表来描述。这些图表可以是流程图、清单、等级图或者任何一种描述项目理论主要元素和关系的形式。这些图表不仅描述了项目理论,而且使项目理论更加集中和具体,以便吸引项目执行者和项目各方的参与。

确证项目理论

项目理论更多地反映了对项目的想象而不是项目的实际运作,项目管理者和政策制定者往往把他们头脑中的项目作为项目的实际,而这样的想象却并不能反映项目的实际情况。项目理论与项目的日常实际相脱离,而人们可能并不明了这个缺点,甚至还使用那些并不符合项目实际的想象来描述项目。

项目理论和项目实际之间的差异是很平常的。实际上,对于这些差异的性质和大小的考察是项目过程评估所要完成的任务,我们会在以后的章节中具体讨论。但是如果项目理论所阐述的项目活动和项目所要获得的结果是实际的项目资源所无法达成的,那么这个理论就过分夸大,也不能起到构造项目理论所设想的作用。举例来说,假如一个职业培训计划要求对象和负责机构之间每个月都要有一次联系,而项目本身却没有职员,那么这部分项目理论就是凭空的想象,因此必须根据项目实际所能达到结果来构造项目理论。

确证项目理论,主要是证明项目人员和项目方是否认为该理论有效地描述了项目运行的机制。如果不能建立一个所有项目方都满意的项目理论,主要项目方对于项目到底应该做什么以及为什么这样做存在着分歧,或者项目包含了互相矛盾的价值理念,那么该项目无疑是不明确的。在这样的情况下,评估者最好以顾问的身份为项目界定一个所有项目方都能接受的项目理论。

对于评估者来说,项目理论的陈述应该是完整和细致的,这样便于进一步的分析和评价。需要注意的是,项目方对所陈述的项目理论的赞同只意味着这个陈述反映了他们对项目机制的理解。而这并不一定表示项目理论是好的。判断一个项目理论的合理性,不仅要基于其很好地被表述,还要经过仔细的评估。下一部分内容将讨论评价项目理论的步骤。

评价项目理论

对项目理论某些特征的评价往往与对项目绩效和项目影响的评估联系在一

起。尽管如此,除了经典的可评估性评价文献之外,一般很少涉及如何具体地进行可评估性评价,特别是对项目设计本身进行评估时尤其如此。但这并不意味着项目理论的评价不重要或者特殊,只是因为可评估性评价是在一般性判断的基础上以非正式的方式施行的,因此不需要做出过多的解释。如果认为项目理论没有问题,那往往是因为证据的缺乏,或由于判断粗略而凸显了其表面合理性。这种情况往往发生在那些服务和对象有直接联系的项目中。比如,一个为增进困居家中的老年人的营养摄入而向他们运送餐饮的送餐项目,就能很好地说明这个问题。

但是大多数项目所依赖的假设和概念是复杂的,不像通过送餐就可以增进营养这么简单。例如,一个家庭维持项目计划通过个别服务的方式向那些不愿把孩子送入托儿所的父母提供社区帮助。在这个项目中,对于项目到底要达到什么目标,怎样达到这样的目标有着各种各样的假设,因此该项目的项目理论很容易出现错误,相应的,就需要对项目理论进行较为严格的评价。

逐个评价项目理论的每个假设和预期是不大可能的,也是没有意义的。但是,还是有一些重要的检验方法可以用来查看项目理论是否像听起来的那样可信。这部分内容,综合了评估者评价项目理论的步骤和各种方法。

关于社会需求的评价

评价项目理论的最重要框架是建立在需求评估基础之上的。正如第 4 章所讨论的,其分析基础是在整体上把握项目所要解决的社会问题和相关目标群体所需要的服务。如果项目理论不能以合适有效的方式说明项目活动和与项目实际性质及其实际社会环境相关的项目产出,那么即使能得到很好的管理和执行,该项目也是一个低效率的项目。因此,通过评价项目所要满足的目标群体的需求来评价项目理论是最为基础的工作。

在评价项目理论是否成功地阐述了满足社会需要的适当途径时,并不是一次性完成的,事实上,这项工作需要做出一系列的判断。特别是做出批评性评价时,就需要与相关的专家和项目各方合作,拓宽评估者的视野和专业性,从而提高结论的可靠性。参与这种合作的可能是具有相关研究和理论知识的社会科学工作者,具有丰富的管理该项目经验的行政管理者,也可能是那些与目标群体联系紧密的支持群体的代表,当然还包括那些熟悉项目的政策制定者和政策顾问。

无论各类合作小组的性质对评价项目理论有什么样的影响,最重要的一点是合作必须是具体的。当项目理论和社会需要只是用泛泛的方式表述时,人们往往持赞同的态度,而如果深入到细节,情况就不同了。举例来说,为了减少大城市日益增长的青少年犯罪,在某地区实行了禁止 18 岁以下青少年在午夜之后在外游荡的宵禁项目。该项目的项目理论认为,宵禁可以使青少年呆在家中,而如果他们在家里,就不会犯罪。因为该项目所要解决的是青少年犯罪问题,所以项目理论乍看起来是符合社会需要的。

但是,经过对细节问题的判断和对项目服务的评估,我们会发现绝大部分青

少年犯罪都是在下午放学之后发生的入室盗窃。除此之外,虽然犯罪者只是青少年群体中极小的一部分,但是他们犯罪的比率高得惊人。进一步说,这些犯罪者往往是那些在放学之后不受监管的"挂钥匙少年"[①]。当我们检查项目理论的细节时,就会发现项目假设主要的青少年犯罪都发生在深夜,而潜在的犯罪者知道宵禁并会遵守禁令,如果他们不服从的话,家长和警察将制止他们。

尽管我们还可以提供更多的细节,但是这些已经足够了,我们可以就此评价项目理论是否与需求吻合,从而找到缺陷。在宵禁项目中,通过研究,可以发现项目理论与要解决的社会问题之间很难联系起来。项目覆盖了整个城市而不是针对人数很少的问题人群,它关注的主要是夜间犯罪,而大多数犯罪却发生在下午放学之后。另外,这些青少年经常触犯比宵禁更严重的法律,他们会乖乖地遵守宵禁法令吗?他们的父母在白天尚且管不住他们,难道他们会在晚上听话吗?同时,警察还要投入相当多的人手逮捕那些违反宵禁令的青少年。经过仔细的讨论,我们越来越怀疑该项目理论的有效性(专栏 5—I 展示了另外一个例子)。

专栏 5—I 用无家可归者的需求作为评价项目理论的基础

第 4 章的专栏 4—J 在说到需求评估时描述了大量无家可归男女的需求。在他们的需求中,占比例最大的是有一个栖身之所、有一份工作、有稳定的收入。将近一半的人(当然也是很大的比例)说他们也需要医药、药物滥用、心理和法律方面的帮助。评估者认为,根据需求所提供的服务就是对这类人群进行干预的指标,这些指标为继续对这类人群的主流群体进行干预提供了基础。因此,出于责任心,项目必须有能力提供和进行这种复杂的服务。

这些发现为项目理论的评价分析提供了两条思路:第一,任何想减轻无家可归问题的项目必须首先解决无家可归者经历的主要问题。也就是说,如果要使无家可归者的条件得到显著改善的话,我们所预期的服务结果(影响理论)必须在大多数有问题的领域展现出一种改善。第二,服务供给体系(过程理论)的设计必须多样化和富有弹性,以期使无家可归者个体在有限资源和困难的条件下能够获得服务。因此,细致地比较任何致力于解决无家可归问题的项目理论的需求评估资料,将揭示进行有效干预的理论重要性。

资料来源:Daniel B. Herman, Elmer L. Struening, and Susan M. Barrow, "Self-Reported Needs for Help Among Homeless Men and Women," *Evaluation and Program Planning*, 1994, 17(3): 249-256.

比较项目理论和各种社会需求的一个有效方法就是分别评价项目影响理论和项目过程理论。每种理论与社会问题联系的方式是不同的,因此,对于项目理论的假设及其要适应的社会环境是否符合的问题就可以十分具体。我们将简要讨论各种理论要素所需比较的要点。

项目影响理论包括项目服务与目标社会环境得到改善的结果之间的因果链。比较项目影响理论和社会需求的关键,是看项目理论预期对社会环境的改善是否符合社会需求。例如,为了使小学生懂得并养成良好的饮食习惯,采取了

① 通常指达到上学年龄,在父母上班时间总有一段时间独自在家、无人照看的少年。——译者注

一个以学校为依托的教育计划。项目所要解决的问题是改善学龄儿童的营养摄入,特别是那些贫困地区的儿童。项目影响理论给出了这样的逻辑:增加教育可以使儿童认识到食品的营养价值,从而导致健康的选择和营养的改善。

现在,假设整体需求评估指出,虽然孩子们的饮食习惯不好,但并不是因为他们缺乏有关营养的知识。需求评估进一步显示,无论家庭还是学校食堂提供的食物都没有考虑营养问题。在这种背景下,项目影响理论就有明显的缺陷。尽管项目成功地告知孩子们更多的营养知识,但是因为他们并没有选择食物摄入的决定权,即使项目取得了自己所设想的最接近的成就,但是项目所要解决的社会问题却并没有得到解决。

另一方面,项目过程理论是一些对项目为目标群体提供容易获得并符合他们需求的服务能力方面的假设。通过评价目标群体获得服务的机会以及他们在接受服务时可能受到的阻碍,我们可以评价项目假设与需求之间的关系。例如,在一个成年人扫盲项目中,设计于晚间在本地中学向干预对象授课。项目过程理论突出了教育和宣传功能,而且向对象群体提供可供选择的多种授课进度。这个框架的细节可能就等同于需求评估数据,从逻辑上和心理上显示了目标群体最需要从项目中获得的支持。例如,孩子的接送问题可能是一个不能不涉及的重要方面,如果不能提供比宣传中更多的个人鼓励,这些成年文盲可能仍不愿意参加扫盲;教师在学识和人格上的吸引力可能是吸引和维系他们参加扫盲的重要因素。通过研究项目理论如何满足这样多维的需求,我们可以评价项目的过程理论。

逻辑性和适当性的评价

全面阐释项目理论应该揭示项目设计本来就存在的关键假设和期望。评价的最有效途径是对项目各组成部分的逻辑性和适当性进行仔细的审查。熟悉项目理论评价的评估者通常建议为评估理论组建一个研究小组(Chen,1990;Rutman,1980;Smith,1989;Wholey,1994)。可以肯定,一个专家研究小组既包括评估者,也包括项目职员和主要项目方的代表。但是,项目各方与项目总有直接的利害关系。要调和评估的立场并增加评估的专业性,吸收那些与项目没有直接利害关系的知情人参与评估是可行的。这些外部的专家可能是相似项目的有经验的管理者,或者是具有相关专业知识的研究者以及支持群体和委托组织的代表。

研究项目理论的逻辑性和合理性是一个相对无结构的和开放的过程。尽管如此,这些研究还是可以找到一些一般性的问题作为评估的参考。以下是研究者可以问到的问题(更详细的资料可以参见 Rutman, 1980; Smith, 1989; Wholey, 1994;也可以见专栏 5—J 的案例)。

专栏5—1 评估马里兰4—H项目:项目理论的明确性和可行性

对马里兰州的4—H项目的可评估性评价是在大量项目文件和对96位项目方成员的深入访谈基础上进行的。有关项目理论的结果如下:

问题:项目的任务和目标是否明确?

结论:4—H项目的整体任务不大明确,另外,项目各方之间、直接卷入项目的执行者和没有卷入的人士之间存在着分歧,对项目任务有不同的说法,如,"向青年介绍农户的生活","增进青年对农业和家庭经济的责任感","增进青年的生存技能"。

问题:项目要影响的人群是否明确?谁是项目的受众?

结论:谁是项目的受众?项目各方和项目职员没有达成共识,书面文件把受众界定为青少年和成年人,8到18岁的青少年通常是项目的传统受众,但是最近也把6岁到7岁的儿童作为受众,一些项目方还认为那些帮助完成项目的成年志愿者也是受众。

问题:在预期效果方面,是否达成了共识?

结论:在州一级的文件里,项目的目标是增进青少年的精神、体质和社会适应能力。在绝大多数项目文件中,都提到了社会适应能力,如自信、领导能力和责任感。但是,很少提到项目可以增进精神素质,更没有其他文件提到项目可以增进体质。

问题:对于预期的效果来说,项目活动是否是可行的?

结论:即使项目所计划的活动都得到了很好的完成,项目活动也不大可能带来预期的效果。项目逻辑缺少一个提供某种课程的环节,也就是说,由于缺少了这样的课程设置,项目职员不知道招募怎样的活动领袖和志愿者,也不知道招收这些人做些什么,这样就妨碍了项目活动的适当性。

资料来源:Midge F. Smith, *Evaluability Assessment: A Practical Approach* (Norwell, MA: Kluwer, 1989), p. 91.

- 项目的目标是否界定明确。项目所要获得的结果应该得到清晰明确的陈述,以便判定项目是否达到了预期的效果。如果满足这个要求,项目目标都是可以观察的,那么确定项目的成功就有了明确的标准和指标。在这个意义上,"引导学生学习计算机技术"不能算好的项目目标,但是"增加学生关于如何使用计算机的知识"则是一个明确和可量度的项目目标。
- 项目的目标是否合理,项目目标是否真的可以被认定是项目活动设想的结果?项目理论应该具体指出那些通过努力、项目本身确实可以取得的预期结果,而不是夸大项目或提供不切实际的过高预期。除此以外,项目理论提出的目标应该是对社会环境的某些方面做出的切实影响,而不是泛泛的意义。例如,"消灭贫困"就是夸大其词。尽管"降低失业率"并不夸大,但是在经济普遍萧条的情况下,这种目标也是不切实际的。
- 项目所设想的改进程序是不是合理?项目所设想的给目标群体带来的收益将依赖于其设计的因果关系链条是否发生,这个链条以项目设想的互动开始,以项目所预期的目标群体的社会环境得到改善结束(项目影响理论)。这个因果链条的每一个环节都至少应该是合理的。项目影响理

论的适当与否决定了项目是否能够产生预期的影响。如果有证据表明设想的因果关系确实发生了,那么影响理论就是完美的。例如,一个项目设想如果利用文献告知那些长期吸食海洛因的吸毒者吸毒有害健康,吸毒者就会戒毒。这种设想根本不是解决问题的方法,也不可能被实地调查结果支持。

- 项目用来界定目标群体、提供服务并在项目完成之前维系服务的步骤是不是清楚和足够?项目理论应该涵盖具体的步骤和功能,以便确定潜在的服务对象和他们的口味并提供相应的服务,更重要的是处理其间的突发事件。除此之外,从项目提供设计服务的能力和目标群体被吸引的可能性角度考虑,这些步骤也是十分必要和正确的。例如,如果要对贫困的老年人进行高血压检查,就应该问一下这样的检查是不是在老人们都方便的地方进行,是不是有比较有效的途径为那些没有固定住所的老人提供服务的场所。如果没有这些考虑,目标群体的大多数人可能接受不了项目提供的服务。
- 项目要素、项目活动和项目功能是不是明确和充分?项目的结构和程序应该是具体的,这样,项目就可以有规范地实施和高效地管理,并且通过有效的指标得到督导。更为重要的是,未达到项目预期的目标、项目要素和项目活动应该是充分和适当的。如果没有人来安排或者根本就没有共识,那么诸如"委托人支持"这样的功能就没有实际的意义。
- 分配到项目及其不同组成部分和项目活动的资源是不是足够?项目资源包括资金,当然还有人力、物力、装备、便利、关系、声誉以及其他成分。项目理论所描述的项目设想应该符合预计(或已知)的、可能得到的用来执行项目的资源。如果项目理论所要求的活动和产出都是项目资源所不能提供的,那么该项目理论就是失败的。如果要按照预期施行项目并且希望得到预期的效果,就应该以现有的项目资源为尺度约束项目理论设想的运作方式和结果。例如,由于缺乏职员,项目培训中甚至连一些简短的活动都不能举行;当然,更不可能希望在培训中获得的管理技能能够产生什么样的显著影响。

通过研究和经验的比较进行评价

尽管在某些方面各个项目都是互不相同的,但是对于如何产生改进、提供服务和实现功能的设想往往有共同之处。由此,在社会科学和人类服务的研究文献中,可能存在着适于评估项目理论不同组成成分的信息。因此,当项目理论形成之后,评价该理论的最有效方法就是看是否符合其他研究结果和实践经验(专栏 5—K 总结了使用这种方法的例子)。

有很多途径可以比较项目理论研究及实践的发现。最直接的方式是检查建立在相似设想基础上的项目评估。这就为判断项目成功的可能性提供某种标尺,或者明确项目存在的问题。在这一点上,相似项目的评估将具有极大的参考价值。

专栏 5—K GREAT 项目理论与犯罪学研究的一致

1991 年亚利桑纳州凤凰城的警察局和当地的教育专家发起了一个旨在防止青少年加入帮派的项目,即 GREAT(预防拉帮结派的教育和训练)。项目获得了联邦政府的资助,并且在全国范围内推广。该项目安排有正式警察对 7 年级的学生进行持续 9 周的训练,项目对课程做了周密的安排,计划向青少年讲述如何确定自己的目标,如何应对群体的压力,如何解决冲突,以及团伙对人生可能造成的影响。

除了格莱泽(Glasser, 1975)的现实疗法(reality therapy)外,没有其他正式的项目理论,但是 GREAT 的训练官员和其他合作者在训练项目督导时参照了社会学和心理学的概念。作为分析项目影响理论的一部分,犯罪学专家提出了关于参加团伙的两个著名的犯罪学理论:戈特弗德森和赫斯奇(Gottfredson and Hischid)的自我控制理论(SCT),阿克思(Akers)的学习理论(SLT)。接着他们对 GREAT 安排的课程进行了评估,评价课程与这些理论的一致性。以下是研究者对第四课的分析结果:

> 第四课　冲突的解决。学生应该学到如何营造理解的环境,这样有助于各方面澄清问题并着手解决问题。本课包含了自我控制理论的相关概念和应对策略,同时,督导员也应用学习理论提供了解决冲突的和平方式。这样一来,就告知学生如何用温和的手段而不是用暴力来解决加入团伙或者与群体意见不一致的情况。这些看法都得益于对团伙和社会学习理论的研究。

通过类似的比较可以看出,项目的课程设置与犯罪学理论之间有着很好的契合,也就是说,GREAT 的课程设置涵盖了自我控制理论和学习理论的内容。

资料来源:L. Thomas Winfree, Jr., Finn-Aage Esbensen, and D. Wayne Osgood, "Evaluating a School-Based Gang-Prevention Program: A Theoretical Perspective," *Evaluation Review*, 1996, 20(2): 181-203.

举例来说,在一个大城市里为鼓励妇女进行乳房 X 射线检查以便及早发现乳腺癌的大众宣传活动中,项目影响理论设想通过电视、电台与报纸的宣传和鼓励,进行乳房检查的比率将会升高。不管影响理论如何验证宣传和检查率升高之间的联系,只要在其他城市的宣传导致了乳房检查率的提升,就说明影响理论是合理的。不仅如此,如果对其他城市宣传活动的评估资料显示项目功能和项目提供服务的计划与项目预期相符,那么该项目的过程理论也得到了某种支持。但是,假定在其他城市并没有关于媒体宣传导致乳房检查率提高的评估结果,那么就可以看其他类似的媒体宣传的结果,如关于免疫、牙检等类似的相关健康检测。因为这些检测的原理相似,任何一个的成功都可以说明以媒体宣传为基础的乳房检查促进活动的项目理论。

在某些案例中,对项目的社会和心理过程所作的基础研究可能会成为项目理论特别是影响理论的框架。从评估的角度看,忽视基础性的社会学研究是不恰当的,这些研究往往对一些重要和常见的干预活动有深入的讨论,研究结果对项目评估大有裨益。例如,仍以上文提到的为鼓励妇女作乳腺检查而进行的宣传活动为例,其内在假设是,劝诫会改变妇女的态度和行为。从更高的层次看,

社会心理学关于态度改变与行为改变之间关系的基础研究可能会给所有媒体宣传项目的影响理论提供一个参照的基础。而已有的研究结果表明,这些宣传信息往往增加了恐惧,而对是否采取行动很少有积极的影响。由此看来,项目影响理论认为只要增加对乳腺癌危害性的认识就可以提高乳腺检查的比率,这种假设是有问题的。

我们还可以找到很多有关媒体宣传的研究,当然还有关于广告和市场的研究。尽管这些文献是有关如何销售产品和推广服务的方法,但是他们仍为评估乳腺癌项目的项目理论提供了广泛的基础。例如,市场细分的研究可能会指出,对于不同人口特征的妇女应该分别在不同时间使用不同媒体进行宣传。这些信息有助于检查媒体是不是最有效的对那些最易患乳腺癌的妇女做出了宣传。

不过,在评价项目理论是否应用其他的相关研究结果和实践总结的文献时,并不仅仅局限在那些与待估项目或程序十分相近的范畴,事实上,往往并不存在对相近项目或相近程序的评估研究。在这种情况下,可以把项目理论分割,然后把与每个组成部分相关的研究结果合起来。项目理论往往可以分成一系列假设条件命题。如果安排个案管理人员,那么将会提供更多的服务;如果学校的教学质量改进了,那么不良行为将会减少;如果提高师生比,那么学生可能会得到更多的重视和关怀。这些命题往往是项目理论的基础,而我们常常可以找到帮助我们评价这些命题的研究文献。对这些命题的评价结果,使我们进一步明确了原来可能不太明显的因果关系,这样就给项目理论的评估提供更广阔的基础(Cordray, 1993; U. S. General Accouting Office, 1990)。

通过初步观察进行评价

项目理论是一种内在的概念体系,不能对它作直接观察。但是,项目理论包括了很多关于项目到底如何运作的设想,因此评估者可以通过观察项目的运作过程,与项目职员和服务对象进行座谈以及其他方式来评价项目理论。实际上,对于项目理论的完整评估,应该包含对项目及其环境的直接观察,而不是完全依赖逻辑上的分析或闭门造车式的评论。直接观察为评估项目理论与项目的一致性提供了实际的检验。

例如,一个项目的项目理论认为,如果向老年活动中心的老人分发有关注意营养饮食的手册,那么65岁以上的老年人对饮食的态度和饮食习惯将会发生改变。观察发现,到活动中心的老年人很少阅读这样的手册,那么项目的关键假设就应该受到质疑。观察结果指出,分发的资料并没有对老年人起到预期的宣传效果,而这却是他们改变态度和行为的前提。

如果要对项目影响理论做出评估,那么观察和访谈的重点应该放在项目预期的产出以及为获得结果而设计的项目服务与目标群体之间的互动过程上。这就要求观察项目预期的产出是否符合项目所处的环境,目标是不是确实能够达到。例如,一个再就业项目设想可以使大部分失业者找到并维持工作,那么评估该项目的项目理论时,就要注意调查当地的劳动力市场,失业者的工作意向(包

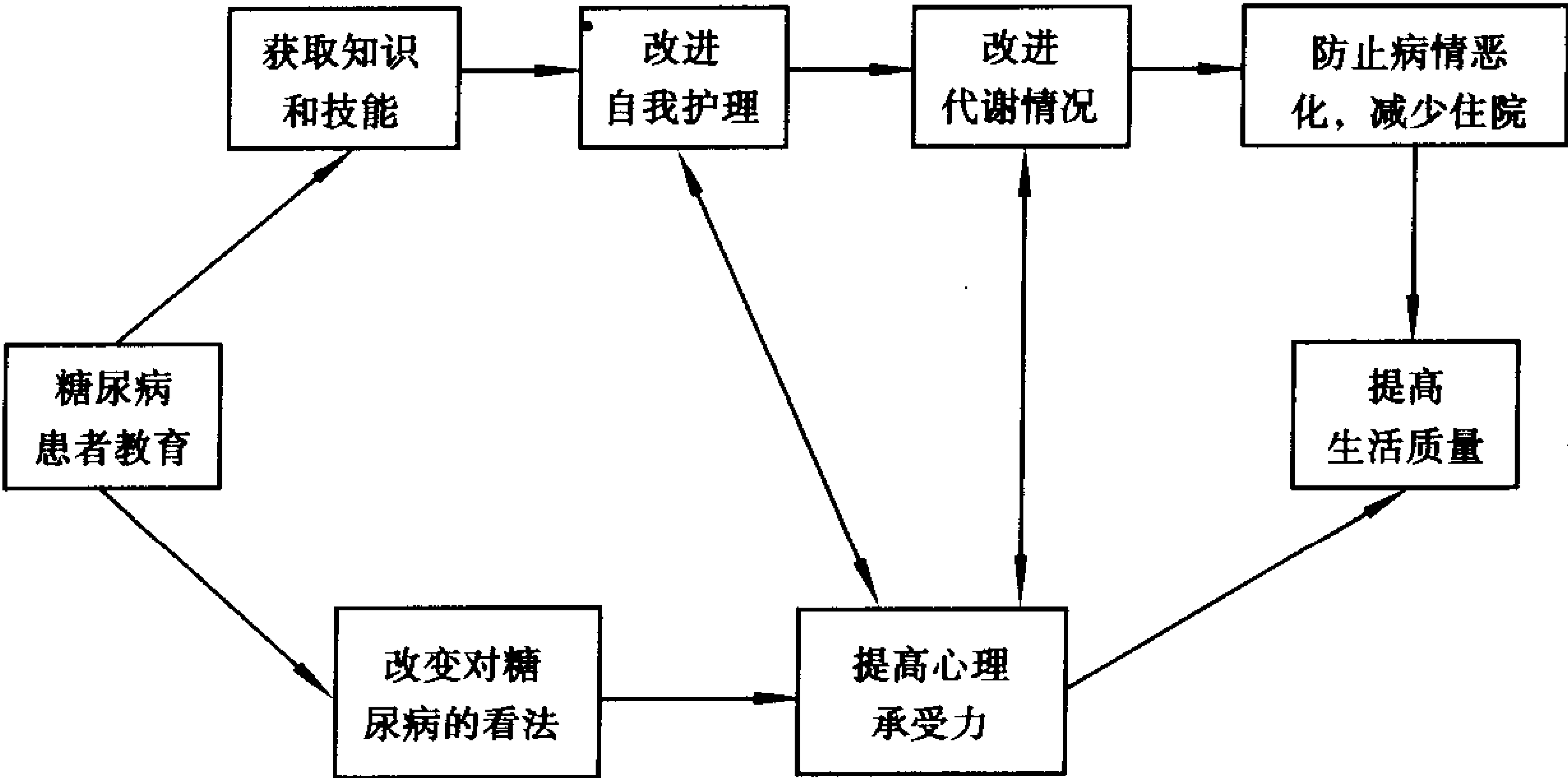
括心理承受能力、技术水平、职业经历和工作动机)和相关工作的经济收益,这样才能评价项目产出是否符合实际情况。在项目提供的服务实施之后,可以通过对职业培训的观察和对参与项目的失业者的访谈,来评估项目活动是否带来了预期的变化。

如果要对项目过程理论的服务送达计划进行评估,就应该观察项目的目标群体,理解他们是怎样、又为什么参与项目,了解他们一旦参与项目,又如何获得服务。这些调查将帮助我们评价项目服务计划是否很好地定位了目标群体,人们是否积极地参与项目并得到了切实的服务。举例来说,如果要对一个旨在减少青少年不良行为而实施的夜间篮球项目进行评估,评估者应该观察项目活动,采访项目职员、参与者和附近的青少年,了解是哪些人参与项目活动以及参与的频率。如果有迹象表明大多数有不良行为的青少年经常参与活动,那么项目的服务计划将得到支持。

项目过程理论的组织计划评估有赖于对项目资源和项目活动的观察。关键是看项目是否能完成预期的功能。例如,一个项目计划要求所有六年级的自然课老师每年向学生提供两次郊游的机会,评估者可以就时间安排、接送以及经费等重要问题与相关的教师和学生进行广泛的讨论,这样就完成了对项目假设的调查。

专栏5—1 检验糖尿病患者自我护理教育的项目模型

糖尿病的日常护理包括以下因素的相互作用:代谢指标、患者自我护理行为以及对致病的社会及心理因素的适应。因此,在糖尿病的治疗过程中,病人是否具有相应的自我照料技能和知识是一个很重要的因素。一些对此感兴趣的研究者特别关注糖尿病对患者本人的影响和意义。他们勾勒了教育患者的影响理论,如下图:



研究者从各州的诊所中获得了220名糖尿病患者样本,通过对调查数据的分析来验证以上模型的关键假设。数据分析应用结构方程模型,当然,这一分析模式仅类似于结构方程模型。变量"对糖尿病的看法"和变量"心理承受能力"之间存在着高度的相关,"知识和技能"与"自我护理行为"之间也高度相关,但是其他变量之间的相关性很微弱。研究者因此认为,当调查数据不能与以上的模型吻合时,就说明原假设——患者本人对糖尿病的看法在他们的日常照料和心理适应过程中起

重要的作用。

资料来源:George A. Nowacek, Patrick M. O'Malley, Robert A. Anderson, and Fredrick E. Richards, "Testing a Model of Diabetes Self-Care Management: A Causal Model Analysis With LISREL," *Evaluation & the Health Professions*, 1990, 13(3): 298-314.

任何涵盖搜集新资料的项目理论评估能比较容易地转换为这个问题,即项目进展是否符合理论预期。实际上,理论检验的经验研究往往是作为评估项目理论的一种精确有效途径而被重点阐述的(参见 Bickman, 1990;专栏 5—L 就是一个例子)。但是,本章主要关注的是被描述为计划的项目理论的合理性问题。也就是说,我们只将其作为项目的预期,而不是对实际情况的描述。因此,在认识到观察和访谈在这个过程中扮演的角色之后,我们并不提倡把对项目理论的评估作为整个项目评估过程中一个十分特殊的评估。相反,我们希望在恰当地描述项目活动、目标群体以及相关的状况和信息之间建立联系,为评估者提供判断项目理论现实性和适当性的有效信息。

项目理论评估的后果和影响

如果一个项目的概念是错误或无效的,那么即使项目的设想得到了很好的贯彻,也不可能取得成功。因此,如果项目理论本身是不合理的,那么也没有正当的理由评估项目的完成、绩效或效率。在可评估性评价的框架下,如果发现项目理论是含混或错误的,那就意味着项目不值得评估。

如果在对项目理论的评估中发现其存在缺陷,那么,正确的反应应该是,对项目进行重新设计。项目理论重构的步骤包括:①明确项目目标。②去掉项目理论中预期不会发生、不必要和不合理的组成成分。③与项目各方合作,获得有关项目目标及项目活动与项目产出之间逻辑联系的共识,评估者在此可以扮演顾问的角色。

如果在评估项目影响和项目绩效时,没有一个清晰可信的项目理论作为依据,就会造成很多含混,而且这种含混是双重的。首先,如果项目过程理论不明确,项目设想的工作是否很好地完成了就是含混的。因为缺乏一个判断工作是否完成的明确标准。所以对各种主要的项目功能来说,就必须通过某种方式建立这样的标准。例如,制订行政管理标准以度量接受服务的对象数量和提供的服务的数量,这些并不包括在整体的项目方案之内。

其次,如果没有项目影响理论作指导,就不能解释为什么发生了这些后果(见第 7—10 章),更为重要的是揭示不了为什么没有发生这些结果。不具体的影响理论限制了评估确定或测量项目后果所依赖的控制变量,相应的,也很难判定为达到设想目标的那些做法的对错。如果项目过程理论也是不具体的,就无法确定项目是不是取得了预期的结果。在这种情况下的评估被称作"**黑箱评估**

(Black box evaluation)",它说明不了项目产出是因为什么而产生。

如果项目理论是明确合理的,就会明确地界定项目的功能和影响,阐述项目设想的活动以及作为活动结果的预期结果。这个结构为评估者和管理者提供了一个比较项目实际进程的、很有意义的基准。依此,项目理论的框架不仅为有效的管理提供了蓝本,而且为评估者提供了评估项目程序设计、项目影响和绩效的指南,这些内容将在随后几章中展开讨论。

小 结

- 项目理论是项目的一个组成部分,可以根据自身情况进行评估。这种评估具有重要的意义,因为如果项目建立在含混或错误的概念基础上,那么获取预期结果的希望就很小。
- 可评估性评价包括描述项目的和目标,评价项目的概念化是否充分并适于评估,确定项目各方对评估结果的兴趣所在。可评估性评价可能导致项目管理者更好地概念化他们的项目。也可能意味着项目没有明确的界定,不适合评估,或者评估的结果也很难得到充分的利用。不过另一种可能是,项目理论非常精细,而且评估结果会得到充分利用,这样的评估就会极有意义。
- 如果要评估项目理论,评估者必须很好地阐述项目理论,也就是说,用项目各方可接受的明确方式陈述项目理论。这样做的目的是,按照预期或者理性构想而非实际状况来描述项目。在此种表述中应该包括三个核心要素:项目影响理论,服务利用计划及项目组织计划。
- 如果构成项目理论的假设组织完好、表述明确,我们就称之为明细化的项目理论;如果它们隐含于项目之中,没有得到完整的表达,我们则称之为隐含的项目理论。当项目理论比较含蓄时,评估者必须在分析和评估前,通过一定的方法提炼并描述项目理论。这些方法包括校对整理项目文件、访谈项目人员和其他项目方及观测项目行动。特别重要的是,要清晰地组织和具体地陈述项目目标和项目活动如何带来预期后果。评估者与项目方的合作能准确而深刻地描述"预期的项目"。
- 有很多方法可用来评估项目理论。评估者对项目理论可做的最重要的评估依赖于对项目提出的具体干预与项目预期满足的社会需求的比较。仔细检查项目概念的细节及其想要解决的社会问题,就会发现项目是否提供了一个改善目标问题的合理计划。当需求评估判断了社会问题的状况时,以上的分析就会得到简化(见第4章)。
- 评估项目理论的一个补充方法是通过项目方和其他知情者评判项目理论是否明确可行、建构恰当,这项工作常常还可以通过评估者对项目理论关键假设的直接观察完成。
- 项目理论的评估也可以应用与项目关键假设有关的其他研究或项目实践的文献。有时可以找到相似的项目,或者是有着相似理论基础的项目,这样,就可用项目理论与相关依据进行整体比较。但是如果研究资料和实践资料不支持整体比较,仍然可以找到与项目假设的具体关系有关的参照依据。
- 评估者也常把直接观察作为其他评估手段的有效补充,以进一步探究项目理论的假设。
- 项目理论评估的结论可能是,项目是没有价值的,其理论基础是有缺陷的。这些结论是对项目理论本身的重要评价,对于项目各方有很高的参考价值。碰到此类情况,一个恰当的回应方法就是重新设计项目,让评估者充当顾问。进一步说,合理的项目理论为评估项目理论的贯彻情况、项目产出以及项目的绩效提供了基础。而这是以后几章内容所要讨论的主题。

基本概念

明细化的项目理论(Articulated program theory):由项目的有关文件直接表述和界定的或者由项目各方和项目评估者共同建构的项目理论。

黑箱评估(Black box evaluation):只针对项目带来的产出评估,而不管项目理论预期的项目产出以及两类产出差异的原因。

可评估性评价(Evaluability assessment):项目评估者、评估主办方和其他项目各方通过调查和交换意见,确定项目评估是否值得进行,以及在评估设计中如何最有效地利用这些评估结果。

影响理论(Impact theory):描述作为诱因的项目活动和作为产出的社会收益之间的因果联系链。

隐含的项目理论(Implicit program theory):没有被完全表述和记录下来,但是项目服务和项目实践中体现出的内在假设和预期。

组织计划(Organizational plan):对项目通过何种方式连接目标人群和项目意图给社会环境带来预期变化的设想与预期。项目的组织计划建立在项目管理视角之上,包含的内容有:项目功能和项目将要促成的活动,项目在执行过程中所需要的人力、财力和物力资源。

过程理论(Process theory):对项目应该如何执行的假设和期望的全面描述,包括项目组织计划和服务利用计划。

服务利用计划(Service utilization plan):对目标群体与项目如何进行初次接触,以及如何在目标人群中完成既定服务的设想和期望。服务利用计划最简单的形式,就是描述指定客户与指定服务在互动过程中相关事件的后果。

6 督导项目的过程和绩效

为了有效地实现预期的社会状况改善，一个项目不仅仅需要是一个好计划，最为重要的是，项目还必须执行其计划，即必须实质性地以预期的方法实现其预期的功能。

尽管项目理念的执行看来似乎很直接，但在实践中通常很困难。社会项目必定包括许多负面影响，这些影响还会危及预期努力达成的项目目标。这样，在项目预期和实际操作之间很容易出现实质性的矛盾。

项目执行是项目过程的具体形式。因此，评估的一个重要功能就是评估项目的执行：日常项目运作中实际发生的项目活动和实际送达的服务。本章将介绍评估者调查这些问题的程序和方法。

签署新法案之后，肯尼迪总统荣耀地对他的助手说："既然这个法案是本国的法律，那么就让我们期待我们的政府来执行吧。"那些上层人士和站在前沿的人士通常都会怀疑社会项目能否恰当地运作。从理念到实际操作，需要许多步骤，并且需要付出巨大的努力才能保证项目与先前的设计和目标相符。这样，任何项目是否得到充分执行，就如评估主办方和管理者所想象的那样，总是一个有争论的问题。

因此，一种重要而有用的评估形式，就是尽力描述一个项目是如何实际运作的，并评估其预期功能执行得如何，这就是我们所熟知的"项目过程评估"，它有个广泛使用的称谓，即"执行评估"。评估的这种形式并不是一种独一无二的评估程序，而是一类方法、概念和将其应用于不同情况和目的的途径。这一评估形式的有限主题侧重于既定项目本身——运作、活动、功能、执行、组成部分、资源，等等。对这类评估方法没有广泛接受的标准，由于主要涉及测量和记录项目运作的信息，我们可以将其称之为项目督导。

什么是项目过程的评估和督导

评估者通常要区分过程（或执行）评估和产出（或影响）评估。过程评估，按照谢尔（Scheirer, 1994）的解释，就是"证明项目是什么和是否按照预期被送达给既定的接受者"。然而，过程评估并不试图评价项目对接受者的效果——那是影响评估的范畴，这一点我们将在随后的几章探讨。

当项目评估是一种持续活动、要进行长时期反复测量时，那就是项目督导。与过程评估和结果评估的区别相对应，**项目过程督导**（Program process monitoring）系统地、持续地记录项目绩效的主要内容。这些记录用来评价一个项目的执行是否符合预设的、恰当的标准。而**产出督导**（Outcome monitoring）是对项目预期产出的持续测量，通常针对的是那些期望改善的社会状况的变动情况。在本书的后面章节，将会讨论与影响评估相关联的产出督导问题。

项目督导通常包括对服务利用领域内的项目绩效评估。督导服务利用涉及对预期目标人群接受预期服务程度的检查。督导项目组织需要比较项目应该做的内容，尤其是计划提供的服务和实际完成的服务的比较。通常有两个关键问题或者其中的至少一个会指向项目督导：①是否找到了恰当的目标人群；②服务是否送达并实现了与项目设计一致的具体功能，或者满足了其他相近类型的衡量标准。项目督导也可以核查现有的或曾有的资源在项目运作中的消耗。

更为特别的是还要督导规划的设计，以用于回答如下评估问题：

- 多少人正在接受服务？
- 那些正在接受服务的人是预期的目标人群吗？
- 他们正在接受服务的数量、类型和质量如何？
- 是否存在没有接受服务的目标人群或接受服务但又不能代表目标群体的

亚群体?
- 目标人群意识到项目了吗?
- 必须的项目功能得到充分执行了吗?
- 项目人员的数量和与必须执行的功能相当的能力都足够吗?
- 项目组织得好吗? 人员能很好地彼此合作、一起工作吗?
- 项目与其他项目、必须与之互动的机构能有效地协调吗?
- 项目资源、人员和基金足够维持重要的项目功能吗?
- 项目资源得到了有效的和充分的使用吗?
- 项目与其管理委员会、基金机构和高层管理者设置的要求一致吗?
- 项目与生效的专业和法律标准相符吗?
- 某些项目场所或现场的项目绩效是否明显比其他要好或坏?
- 参与者对其与项目人员之间的互动及互动程序和方式满意吗?
- 参与者对其接受的服务满意吗?
- 参与者在接受服务之后具有恰当的行为吗?

设置项目过程的评判标准

认识上述项目督导问题中的这些评估主题尤其重要。在所有问题中,关键的术语有:恰当的、足够的、充分的、满意的、合理的、所希望的和其他表示评估判断必要性的词语。因此,要回答这些问题,评估人员和其他责任群体不仅必须描述项目绩效,而且要估计其是否令人满意。反过来,这就要求做一些基本的判断,即应用一些正当的标准或规格。在这些标准还没有清楚地说明和认可的情况下,评估者会发现建立可操作的标准与确定项目在相关维度上的绩效一样困难。

设置项目绩效标准的方法有很多种。而且,不同的方法只能用于不同的项目绩效维度,因为对适当项目对象数量的界定不同于对充足项目资源的界定。这就是说,在项目督导中,具有最广阔范围和最一般效用的标准问题,就是第5章中描述的项目理论的应用。

正如我们展示的,项目理论被划分为项目过程理论和项目影响理论。项目过程理论就是描述项目要遵循的形式,这些形式实际上构成了项目要做什么或怎样做的计划、蓝图,并因此使得目标群体获得恰当的服务。项目影响理论是描述通过有效服务而预期的结果以及构成结果的因素。进一步讲,这些描述完全建立在需求评估的基础上(是否是系统的或非正式的),并因此把项目设计与要改善的社会条件联系起来。当然,项目过程的设计和采纳通常同时包括投入和项目各方的努力。因此,项目理论在描绘项目"应该"做什么和相应的项目绩效有什么样的构成方面具有某种权威性。

因此,在这一意义上,项目督导就是建立在项目理论尤其是过程理论框架之上的活动。项目过程理论描述了关键的、被假定为对有效的项目是必需的部分、功能和关系等,因为那就是项目的主要目标。这些信息对督导者而言,就是识别

项目绩效的重要方面。然而,作为项目蓝图,过程理论也给出了一些要达到的绩效指标,并因此提供了判断实际绩效是否合格的基础。前一章的专栏 5—E 系统地说明了出院精神病人照料项目的项目过程理论的服务利用部分。为了方便起见,流程图一步一步地展示了与出院病人之间的互动和经历,以及由此获得的项目服务的结果。一个全程督导程序会系统地记录每一步实际发生的事情,报告服务利用的每一个重要方面。因此,在专栏 6—A 中,服务利用的第一个作用就是说明重要的事件,以便能搜集相关的信息。项目督导应以系统的方式编成文件,每一步中都需要如此。例如,在这个照料项目中,某个督导程序应该报告每个月有多少病人从医院出来,社会工作者访问的比例是多少,有多少人求助过服务和求助哪项服务,有多少人实际接受了服务,等等。

如果项目过程中应该发生的事情却没有发生,这就表明项目执行得不好。当然,在实践中,事情不会这么简单。一些重要的事件不一定完全出现或完全不出现,而是会或多或少地出现。就这里的流程图而言,社工不一定访问所有的出院病人,而是访问了部分,另一些出院病人可能求助于其他服务,等等。当然,就服务而言,还有质量维度问题。如果一个出院病人求助于几个社区服务机构,但却没有获得适合他或她需要的服务,这就不能代表好的项目绩效。为确定我们到底要做多少或做得如何,我们需要一个与督导程序信息体系平行的附加标准。那就是,如果督导程序报告说 63% 的出院病人在出院几周后获得了社工的访问,我们就不能说绩效指标没有告诉我们什么百分比是“好的”。如果我们的期望是 100%,63% 是坏的绩效吗?或者在难以确定服务对象的条件下,这是一个令人印象深刻的绩效吗?

在这种情况下,最普通和广泛应用的标准就是简单地使用**管理标准**(Administrative standard),即由项目管理者或其他责任群体设定的、要求达到的标准。例如,一位项目主任和员工可以送达 80% 的服务,或使 60% 的项目参与者在接受项目工作培训后持续工作 6 个月。对出院照料项目来说,其管理目标可能是使 75% 的病人在出院两周后获得访问。这样,从项目督导中得到的 63%,就说明项目绩效不合格,尽管不是特别低于标准。

项目绩效的管理标准和目标可能根据经验、相关项目绩效或仅仅根据项目管理者或倡导者的专业判断而设置。然而,如果合理地加以证明,这些标准就能提供有意义的、用以测量项目绩效的评估标准。在相关特性中,项目绩效的某些方面也许并不符合已有的法律、道德或专业标准。譬如,上述“照料标准”在治疗一般疾病的医学实践中已经被采纳,这样就提供一套进行医疗照料项目绩效评估的实践性标准。相似地,儿童保护服务项目依法则要提供如何处理可能的儿童虐待或忽视的信息。

因此,在实践中也必须重视这样的事实:项目绩效特殊维度的评估通常没有具体的、预先确定的标准,而往往是一种事后评判。这就是“见好才好”的、评估好的项目绩效的信条。就注意禁毒传媒的高危少年而言,搜集项目督导资料的评估者发现,在项目员工和其他主要项目方中,对于什么样的比例是可接受的,

没有一致的观点,他们自己就是模糊的。然而,如果结果是50%,考虑到总体的特性,人们也许会达成某种共识,认为这样的结果不错,尽管有些项目方可能在看到资料之前期待更高的比例。在其他一些研究中,如果结果只有40%或者60%,也能认为是不错的。只有在很极端的结果中,比如只有10%,那么,所有项目方就会对如此低的比例感到忧虑、不满。总之,如果没有具体的前测标准,许多项目绩效就都会被认为是可接受的。当然,如果测量程序和标准过于灵活以至于所有的都能"过关",也是没有用的,评估也就失去了"门槛"的意义。

非常相似的考虑因素也适用于过程理论的组织构成。第5章的专栏5—F展示了出院照料项目的组织计划。回过头来看就会发现:根据恰当的标准,就可以督导和评价项目绩效的可识别维度。例如,在那个计划中,要求个案管理者访问项目对象及其家庭、评估服务需求和做好转诊服务,等等。项目督导程序就应该分门别类地记录项目活动,为评估提供信息。

项目督导的一般形式

在项目评估中,尽管项目绩效的督导和评价有许多相同的地方,但两者使用的方法很不一样,使用的术语也不相同。这类评价也许是一次性的或连续性的,所以,信息的产生和获得会有一个过程。这样的过程可以由外部评估者或项目代理机构雇用的评估者来主导(实际上这是作为一种没有专业评估者参与的管理工具而设置的)。而且,督导的目标可以为管理目标提供反馈,说明主办方和决策者的责任承担能力,提供独立的过程评估或影响评估。有鉴于此,我们区分了项目督导的两种主要形式:过程或执行评估、日常项目督导。

过程或执行评估

典型的过程评估是由评估专家进行的,可能会涉及项目工作人员,但并不被整合到项目日常活动中的独立评估活动。依照惯例,过程评估完成后,一般地会向项目管理者和其他项目方提供项目绩效的信息,但不是日常的、连续的项目操作的一部分。专栏6—A描述了儿童综合服务项目的过程评估。

专栏6—A 针对儿童综合服务的过程评估

许多分析家发现,传统的儿童分类服务基金体系需要按照严格的资格审查等规则,把基金分配给相应的专门类别,以至于没能很好地服务于儿童。有批评者认为,应该把零散的各种服务综合起来,相互协作,这样就能够获得更加有效的服务。1991年约翰逊基金(Robert Wood Johnson Foundation)启动了儿童健康起步(Child Health Initiative)项目,用来检验综合各种儿童服务和资助、进而探讨系统改革的可行性。具体地说,创新项目包括以下组成部分:

- 综合机制,即把各类已有的项目基金综合在一起,建立单一的儿童健康基金。
- 使用案例管理的协调程序,使用综合基金为有需求的儿童提供综合、连续的照料。
- 督导系统,识别社区儿童的健康和相关需求及其和既有服务的差距。

在全国范围内,有九个地方被选中来进行这个项目的示范。加州大学旧金山校区的健康政策

研究所要对这些项目进行评估,并要达到两个目标:①测量项目执行状况与原计划的一致程度(保真模型),②评估每一个项目组成部分的执行程度。第一年,评估工作侧重于政治、组织和项目的设计。第二年,评估工作的重点转向执行和初步产出。综合使用各种方法,包括实地观察、对项目管理者书面调查、主要参与者的深度访谈、服务提供和接受双方的专题小组调查、项目相关文献的调查。

评估发现,在九个示范点中,大部分在监护和协调方面都取得一定程度的成功,但没有一个示范点能够建立综合机制。对每一组成部分的发现如下:

- 综合:有几个示范点成功地创立了一些小的、灵活的基金,但都不是来源于既有的各种项目基金。没有一个示范点能够按照计划进行综合。
- 照料协调:大部分示范点在项目对象层面通过案例管理成功地执行了照料协调,但没有在系统层次上实现协调。
- 督导:各示范点在完成这项任务时都遇到了很多困难,但大部分都获得了一些进展。

资源来源:Claire Brindis, Dana C. Hughes, Neal Halfon, and Paul W. Newacheck,"The Use of Formative Evaluation to Assess Integrated Services for Children," *Evaluation & the Health Professions*, 1998,21(1):66-90.

作为一种评估方法,过程评估扮演着两个主要角色:第一,如果仅仅需要知道项目运作、服务送达和类似项目的绩效,过程评估就可以作为一项独立的评估活动起作用。下面的几种情况与这个阐释是相一致的。对相对较新的项目而言,独立的过程评估是很有意义的,例如,探讨在新项目中如何很好地运作和服务。项目过程通常是正式评估设计的重点,目的是为新项目的管理者和主办方提供有用的反馈。当提出项目是如何被组织的、服务质量如何或者到达目标人群的成功性等问题时,过程评估针对的往往是既有的项目。当项目是要送达已知有效的或假定有效的服务、进而最重要的执行问题是服务是否恰当地送达的时候,过程评估就是项目评估的主要方法。例如,在一个受管理的照料环境中,过程评估可以用于评估在各种诊断中是否遵循了规定的医疗方案。

过程或执行评估通常也和影响评估一起使用。实际上,不包括过程评估的影响评估,通常是不可取的。项目所强调的、对社会条件产生影响的先决条件就是,项目活动被认为能够对需要改善的社会状况产生影响。因为,在很多人类服务领域内,在现有的基础上,维持项目的运作并把恰当的服务送达给目标人群,都会遇到巨大的挑战。所以,一般来讲,想当然地认为项目按计划执行,是不明智的。因此,充分的影响评估一般都需要过程评估的辅助,用以确认项目所提供服务的质量和数量。这就意味着,这些信息能与项目服务影响方面的信息进行综合。

日常项目督导和管理信息系统

项目督导的第二种主要形式,由针对项目过程的某些日常督导指标的评价构成。在社会项目的管理中,通过有规律地提供关于项目执行情况的信息反馈

来督导项目过程的重要指标,是一种有用的工具。这种反馈使管理者在问题出现时能采取正确的行动,也能为项目各方提供项目执行状况的有规律的评估。因此,过程评估的一种形式就是整理社会项目的日常信息系统,以便获得、编整和定期总结回顾相应的资料。这样,过程评估便成为了与人类服务项目共存的**管理信息系统**(Management information system, MIS)。专栏6—B描述了用于婚姻和家庭咨询服务的MIS。

MIS经常提供以下信息:客户对所提供服务的意见、员工提供的服务、诊断或项目参与的理由、社会人口资料、治疗和花费、产出状态,等等。这个系统的一部分是列出项目对象(或基金受众)、讨论服务的花费和储存其他信息,如项目对象的病历、目前参与的其他项目。在许多情况下,MIS能代替过程评估,因为过程评估搜集的大量信息在项目MIS中可以得到。即使当项目的管理信息系统不能满足过程评估的需要,也可提供大量的评估者所需要的信息。管理信息系统所提供的资料,能同时为管理者和评估者使用。

专栏6—B 以色列婚姻家庭咨询综合信息系统

婚姻与家庭咨询机构是由特拉维夫(Tel Aviv)市福利部和特拉维夫大学夏伯尔(Shapell)社会工作学院联合支持下建立的。这个机构提供婚姻和家庭方面的咨询、为本市最贫穷的部分居民——犹太人、穆斯林和基督教徒提供社区服务。

这个机构的综合信息系统是根据向他们寻求帮助的项目对象的要求设计的。希望通过督导过程和产出以及提供组织和临床决策所需的资料来服务于机构和个体咨询者。为此,资料以三种形式搜集,并组织成计算机化的信息系统。资料的内容包括:

- 由项目对象提供的背景资料,例如,社会人口特征、医疗和心理治疗史、寻求帮助的问题、问题的紧迫性、对治疗的期望和找到门诊的途径。
- 麦克马斯特(McMaster)临床评分量表,这是督导者根据家庭运行和家庭健康的六个维度确立的标准化量表;咨询者为每一个项目对象每月填一次表。
- 治疗完成后的回顾性评估,包括两种表格,一种表由咨询者填,另一种由项目对象填。例如,治疗的实际问题,如:持续时间、处理的问题、项目对象和咨询者对问题达成共识的程度、是否有问题没有提出来和为什么没有提出、过程的回顾性评估、既有问题和麦克马斯特功能性领域改善程度的评估、项目对象和咨询者对过程和结果的满意程度。

咨询者在他们想得到资料时就可以进入这个系统找到资料、无论咨询者是想获得资料,还是为诊断寻找依据,都可以获得项目对象近三个月状况的图表。诊断管理报告也由此产生。例如,根据项目对象民族分布,就可能发展出一个阿拉伯社区中心,以便更好地为这个群体服务。管理报告还可以描述治疗结束的方式和时间、项目对象带到机构来的问题以及虽然求助却没有出席第一次会议的人的百分比。信息系统也用于研究目的。例如,治疗成功的预测研究,项目对象和咨询者对治疗过程和结果的比较性理解,以及存在问题中的性别差异。

资料来源:Rivka Savaya,"The Potential and Utilization of an Integrated Information System at a Family and Marriage Counselling Agency in Israel," *Evaluation and Program Planning*, 1998,21(1):11-20.

项目过程督导的各种观点

项目督导的目的是否是出于评估者、项目管理者和员工或政策制定者、主办方和项目各方对信息的需要,这里有、并且也应该有相当大的重叠性。理想地说,督导活动作为评估的一部分,应该满足所有这些群体的信息需求。然而,实际上,时间和资源的限制,可能要求优先考虑某些信息需求而不计其他。过分强调观点的差异性是很危险的,但为了区分差异,描述三个主要“消费群体”对项目督导目的的立场是有益的。尽管有很多特例,但是三个主要的“消费群体”对项目督导的目的在立场上还是有很大的差异,并且,这些差异会影响项目督导的结果。

从评估的立场来看督导

对评估研究者而言,督导项目主要是出于许多实际的考虑。经常发生这样的状况,因为适当的干预没有送达或没有把干预送达给目标群体或二者同时存在,进而使得项目影响激减,有时甚至减到零。据我们估计,许多项目失败就是因为这些执行问题,而不是缺少潜在的有效服务。因此,督导研究,本质上是理解和解释影响因素。知道所发生的事是解释或推测项目运行或不运行的先决条件。没有督导,评估者就在“黑箱”中研究;没有根据地推测更大的项目或不同的干预送达方法是否会改变影响的结果。

从责任的立场来看督导

督导信息对主办和资助项目的人来说也是很重要的。项目管理者有责任对项目主办和资助者告知项目活动、项目绩效的程度、遇到的问题和将要发生的事(这个问题的一个方面参看专栏6—C)。然而,评估者经常被命令提供相同或相似的信息。事实上,项目的主办者和资助者在一些情况下希望评估者成为“他们的耳目”,成为提供项目具体状况的第二个消息来源。

政府和资助团体,包括国会,都处于大众媒体的注视之下。对批准项目的合法团体和充当政府耳目的组织而言,他们的行为也是可见的。例如,在联邦层次上,管理与预算署(The Office of Management and Budget)作为行政机构的一个部门,对项目发展、筹划和花费具有相当的权威性。美国审计总署(GAO)是国会的一个部门,为参众两院的议员提供有关项目效用方面的咨询,有时甚至直接参与评估。州和大城市的政府都有类似的部门。任何接受外部基金的社会项目(无论是公共还是私人的基金的项目),都不可能期望避免详细审查和逃避责任承担。

除了资助者和主办者外,其他项目方也要对项目承担责任。面对纳税人对社会项目支出所持的保留态度,以及由于缩减基金而导致的资源竞争的加剧,所有项目方都在详细审查他们支持和不支持的项目。有关集团利用督导信息来游说,使其倡导的项目扩展或找到其相应的自身利益,同时使其反对的项目缩减或

被遗弃。应当记住的是,项目各方包括了项目对象。项目对象立场的一个戏剧性例子就发生在里根总统打电话给一位人工心脏移植病人、希望他好起来的时候,这时,全国所有的听众听到病人抱怨没有收到他的社会保障金支票。

专栏6—6 项目和服务的利用研究

任何服务组织,尤其是在资源紧缩的时代,都需要对项目服务和活动进行评估。通过评估,能够发展和维持其所需的、适应环境变化的灵活性。这意味着组织需要进行自我评估,即使是在一个理想的世界里。自我评估需要持续地评价自己的活动和目标以及对结果的利用,若有必要,还要修正组织的项目、目标和方向。

在机构里,评估的基本功能是提供目标达成的、项目对主要对象有效用的资料,这些对象包括行政部门、中层管理者和管理委员会。主要的对象,特别是行政部门和委员会,经常面对来自外部环境的重要质询,如立法者和基金会。这些质询通常集中在项目对象对服务的利用、可及性、连续性、丰富性、结果或效用、成本等问题上。信息的建立就是使用模式或项目对象利用模式研究。使用模式研究(无论是简单质询还是详细、复杂的调查),基本上都是描述性。它描述的是谁使用服务和怎样使用服务,并且当与组织的要求或目的相关时,这样的描述就变成了评估。

资料来源:G. Landsberg, "Program Utilization and Service Utilization Studies: A Key Tool for Evaluation," *New Directions for Program Evaluation*, no. 20 (San Francisco: Jossey-Bass, December 1983), pp. 93-103.

很清楚,社会项目是在政治性环境里运行的。由于涉及利益关系,这样的情形几乎没有例外。人类和社会的服务业不仅在资金和雇用人数上巨大,而且还承载着意识形态和情感的包袱。项目通常有畅所欲言的社区成员大军支持或反对;事实上,社会项目部的游说努力比得上国防工业,由此,政治家对特定项目所持的态度通常决定他们在选举中的命运。责任承担力信息是项目各方在支持者与反对者的斗争中使用的主要武器。

从管理的立场来看督导

管理性督导(包括使用 MISs)通常关系到与评估和项目责任承担研究相同的问题,差别在于目的不同。在督导资料中,评估者的兴趣一般集中在项目影响与项目执行相联系的资料上。责任承担研究主要给决策者、主办者和其他项目方提供信息,以便判断项目活动的适切性和决定项目是否应该继续、扩展或者收缩。这些研究可以使用相同的、由项目管理人员建立的资料库,但他们通常用一种批判精神来使用。相反,管理性督导活动与决策性判断联系更少,而更多地涉及作为项目运行规则部分的矫正性测量。

从管理的角度来看督导在项目执行期间特别重要,而对于新项目的探索性检验,尤其是创新性项目来说,亦复如此。无论这些项目计划得如何好,在执行过程的早期,意料之外的结果和负面影响通常会有所暴露。项目设计者和管理者需要迅速、充分地知道这些问题,以便尽可能快地在项目设计中做出改变。例如,假设有一个用来帮助工作母亲的医疗门诊只在白天开放。督导发现,虽然项

目对象对这样的门诊服务有巨大的需求，问题是，门诊时间恰巧使得目标群体中的大部分人失去了就诊的机会。或者假设在儿童中普遍存在的严重心理问题会在学校里表现出来，那么就需要一个项目来应对这样的问题。但是，如果早期发现大部分儿童没有如此严重的心理问题，那么，项目就能由此得以调整。

对于已从探索阶段转向实际运作阶段的项目来说，项目督导通过提供覆盖面（项目到达预定目标群体的范围）和过程方面的信息服务于管理需要，并因此反馈出项目是否满足了规定。在督导信息显示没有达到目标时（即项目绩效的成本高于最初的投入，或项目人员工作负荷太重或太轻），就必须对项目进行调整。忽视项目督导的管理者，就会严重、系统地面临项目明显不同于项目对象需求的困境。

督导信息对管理和评估都是有用的，在此，有一些问题必须要估计到：搜集和报告多少信息是明智的、采用什么形式和频率进行报告、可靠性怎样、评估者和管理者不一致方面问题的机密性如何。例如，对于非赢利性儿童娱乐项目，有经验的管理者也许认为，最重要的信息就是每周的参与情况，因为这个信息可以用来向项目管理委员会描述项目的状况。然而，评估者可能对按月或季度汇总资料感兴趣，但也可能认为，在报告之前，应用天气差异、假期等因素进行校验——甚至这些必要的检验会用到复杂的统计过程。

另一个要注意的就是资料的所有权问题。对管理者来说，项目创新结果方面的督导资料应当保密，直到董事会的研究委员会讨论并提交给董事会。评估者可能希望立即写篇论文发表在《美国评估杂志》（*American Journal of Evaluation*）上。或者，由于在项目对象中遗漏了某个种族群体而导致项目管理者立即更换专业服务主管，而评估者的反应可能是要研究导致遗漏的原因。由于项目人员和评估者之间的各种关系，所以，这些问题的解决是根本性的。

注意：项目管理和实施的许多方面（如遵守税务规则和雇佣法或协商联合契约等）是评估者没有能力来评估的。事实上，受过社会科学训练的评估者，（尤其）是那些开始就在学院工作的人，也许并不适合管理任何项目事务。记住这一点是非常明智的，那就是，即使能够分享来自于 MIS 的资料，也不意味着评估者应该介入组织的管理工作。

在本章的剩余部分，我们将集中讨论在服务利用和组织管理领域内，督导项目过程和项目结果的概念与方法。只有在这些领域里，接受过社会研究训练的人的能力才是有用的。

服务利用的督导

项目过程督导的一个关键问题，就是了解目标人群实际接受项目服务的程度。目标人群参与涉及项目管理者和主办者双方。有效地管理一个项目，需要目标人群的参与保持在一个可接受的水平，并在低于这个水平时能够采取正确

的行动。从项目主办方的观点来说,目标人群的参与是项目有效性和服务需求程度的一个关键标准。

服务利用的督导对项目参与者自愿的或必须学会新程序、改变其习惯或接受指导的介入来说,是尤其关键的。例如,计划提供广泛服务的社区精神健康中心,经常不能吸引那些可以从该服务中获利人群的大部分、从精神医院出来的病人并没有受到利用社区精神健康中心服务的鼓励,也通常不与中心联系(Rossi, Fisher, and Willis, 1986)。相似地,一个计划为准备买房者提供信息的项目发现,很少有买房者寻求这样的服务。因此,项目发起者需要注意的是,如何更好地激发潜在目标人群寻求并参加到项目中来。例如,在某些特殊情况下,目标人群需要付出极大努力才能参加项目或对项目现场的地理位置给予特别的注意(Boruch, Dennis, and Carter-Greer, 1998)。

覆盖面和偏差

服务利用问题可以简单地分解为覆盖面和偏差。**覆盖面**(Coverage)是指与项目最初设计相比,目标人群实际参与的水平;**偏差**(Bias)则是指次级群体的参与比例高于其他群体的程度。很清楚,覆盖面和偏差之间是有联系的。一个包括了所有计划参与者、并没有其他人群加入的项目,其覆盖面方面是没有偏差的。但因为很少有社会项目能够达到这样的程度,所以,偏差就变成了问题。

偏差可能来自自我选择,即一些次级群体比其他群体更频繁地自愿参与;也可能来源于项目活动。例如,某个项目人员可能喜欢一些项目对象而拒绝或阻碍其他人。项目共同面对的一个诱惑是选择最"有成功倾向的"目标人群。由于一个或多个项目方的自我利益,这些"浮云"经常出现(专栏6—D中就描绘了一个典型的例子)。最后,偏差可能来源于一些不曾想到的影响,如由于项目办公室的位置可能会使得临近的次级群体更多地参与项目活动。

尽管有许多社会性的项目比如"食物券"(food stamp)项目,希望能服务于目标群体中的大多数人;但是很显然,项目资源仅能为部分目标群体提供服务。因此,后来在项目策划和发起阶段,对目标群体的界定经常不够具体。项目人员和主办者可以通过更严格地界定目标人群的特征和更有效地使用项目资源来校正这个问题。例如,如果通过建立一个健康中心来为没有医疗照顾资源的人提供医疗服务,就可能导致大面积的需求,以至于许多真正需要这种服务的人得不到服务。解决这个问题的办法就是提高资格标准,即在保证给予最需要服务的人群以服务的基础上,用健康问题的严重性、家庭规模、年龄和收入来衡量和筛选,以减少目标人群的数量。在某些如WIC(妇女婴儿及儿童营养计划)或为穷人提供住房担保的项目中,低覆盖面是一个系统性问题;国会从未提供过足够的资金满足所有的有资格者,对此,只能希望国会能在未来扩大财政预算。

与此相反,超过覆盖面的情况也会发生。例如,电视节目"芝麻街(Sesame Street)"吸引的观众大大超过了最初的目标人群——有缺陷的学前儿童,而包括了健全儿童甚至成人。因为这些额外的观众是在没有额外成本的情况下进入

的，虽然这种超过覆盖面的情况不会耗尽资源，然而又确实不合"芝麻街"项目的最初目标，因为该项目是为缩小缺陷儿童与正常儿童之间的学习差距而设计的。

在其他情况下，如果超过覆盖面，就会产生成本上升和别的问题。例如，由教育部资助的双语项目是为许多主要语言为英语的学生而创立的。由于经费的数量是依据双语班的注册人数来给的，一些学校系统就通过招收不符合预定条件的学生使参加者的数量膨胀。还有的情况是，学校把双语教育作为针对"问题儿童"的班级，这样，在双语班上就充斥着训戒。

专栏6—D 失业的"浮云"

当提供公共服务的管理者为项目对象中占据最有利位置的人群体提供不成比例的服务时，实际上就等于给服务利用的效果掺进了水分。美国就业服务部(the US Employment Service (USES))提供了一个清楚而重要的关于"浮云"的例子，那就是USES扩展、减缩和重新组织，并由此存在了半个世纪的实例。USES认为，项目的主要目标是为雇主提供工人、而不重视为工人提供工作。这使USES向失业者展现了最好的期望，并减少了失望。

USES的管理者(在这项目建立以后产生的)强调以雇主为中心服务的必要性，而不仅仅只是愿望而已，这一点也令人惊奇。根据设计，项目的成功依赖于对雇主的服务、而不是"无条件"的失业者。正如约翰逊总统迫于城市就业压力，大约在1965年沃特(Watts)暴乱前两周写的那样，"我们在赢得和帮助真正的、'无条件的'劣势群体方面，还没有取得任何重大的进步。"

资料来源：David B. Robertson, "Program Implementation Versus Program Design," *Policy Study Review*, 1984, 3: 391-405.

然而，在社会干预中，最通常的覆盖面问题就是没有充分获得目标人群的参与，要么是由于参与者招收或保留方式的偏差，要么是由于项目潜在对象不知道、不能使用或拒绝使用项目。例如，在大部分就业培训项目中，只有少数由于失业而想加入项目的人才具有资格。相似的情形也发生在精神健康、药物滥用和其他项目中(参见专栏6—E)。现在让我们回到这个问题：项目覆盖面如何作为项目督导的一部分来测量。

专栏6—E 为无家可归者提供食物券项目的覆盖面

在美国的城市庇护所和食物供应处(food kitchens)，对100 000无家可归者进行抽样，进行严格的调查，据此，波特(Burt)和科恩(Cohen)实实在在地给出了一些为我们所知的精确维度，包括：无家可归者所获得的、赖以生存的食物在数量和营养方面都是不足的；收入略高于临界点的人群根本就没有办法获得足够的食物。无家可归者不挨饿，就是食物供应处和庇护所无成本地为他们提供饭食的最大成果。

由于从收入方面看，大部分无家可归者都有获得食物券的资格，因此，他们参与项目的比例应该很高。但他们并没有得到好处，波特和科恩报告说，只有18%的调查对象正在接受食物券，其中几乎一半的人从未使用过食物券。这主要是因为获准得到食物券要通过资格审查，即过一些行政性的程序。这对许多无家可归者来说是不容易的，他们不可能有所要求的文件、获得食物券的说辞

或填表的能力。

而且,食物券项目还基于这样的假设:参加者能够在当地食物供应处很容易地获得食物,以备饮食之需,并且有能力储备食物。但这个假设不适用于无家可归者。当然,食物供应处确实出售一些即时食物,而且可以很灵活地配餐。由此,食物供应处能够从无家可归者的食物券中获得一些收益,但是对大部分无家可归者来说,食物券相对无用。

1986 年通过的法律允许无家可归者用食物券交换非赢利组织提供的食物,在庇护所建立提供食物的场所,使得在庇护所居住的人有获得食物券的资格。然而,通过调查食物提供者、庇护所和食物供应处,波特和科恩发现,很少食物提供者申请了接受食物券的资格。样本中的大约 3 000 个食物提供者中,只有 40 个具有合法资格。

而且,在有资格接受食物券的食物提供者中,大多数从未开始或刚开始不久就放弃了接受食物券。这使得以食物券当作食物支付的实践变得毫无意义,因为这些食物是免费提供的,这样,在排队买食物的人中,尽管要求手里有食物券的人要用食物券来支付所得到的饭菜,而提供食物的人却没有因此获得任何收益。能够使用这个系统的食物提供者就只能这样:要求获得食物的支付现金或劳动;对这个项目来说,食物券成了这些支付的替代物。

资料来源:Martha Burt and Barbara Cohen, *Feeding the Homeless: Does the Prepared Meals Provision Help?* Report to Congress on the Prepared Meal Provision, vols. Ⅰ and Ⅱ (Washington, DC: Urban Institute, 1988). Reprinted with permission.

测量和督导覆盖面

项目管理者和资助者同样都需要关心覆盖不足和覆盖过度的情况。低覆盖面是用实际参加项目与需要参加项目的目标人群的比例来测量。覆盖过度有时被表述成不需要项目的项目参与者的数量与参与项目的总人数相比较。项目资源的充分利用要求使需要项目的人员获得服务的数量最大化和不需要项目的人员获得服务的数量最小化。

测量覆盖面的难题就是,几乎总不能确定需要服务的人数,即目标人群的数量。这样的需求评估程序在第 4 章已经进行了描述,如果作为项目计划的一个整体部分来执行,那么通常会将这一问题缩减到最小。另外,有三个信息来源用于评估项目服务与恰当目标人群的程度,这三个来源是:项目记录、项目参加者调查和社区调查。

项目记录

几乎所有的项目都保留着对目标人群服务的记录。来自妥善维护的记录系统(特别是来自 MISs)的资料通常可以用来估计项目覆盖面和项目偏差。例如,关于进入项目的不同筛选标准的信息,可以制成表,以确定获得服务的单位是否是项目设计中所说明的单位。假设某个计划生育项目的目标人群是妇女,她必须小于 50 岁,具有至少六个月社区居民资格和有两个或两个以上不到 10 岁的孩子。项目参加者的记录是可检验的,以检验实际接受服务的妇女是否符合特定的条件、特定年龄或团体的标准,是否低于或高于所给的标准。这样的分析,

也根据合格特征或综合特征排除了项目参与中的偏差。另一个涉及无家可归者的公共庇护所效用的例子呈现在专栏6—F中。

然而,项目记录的质量和范围有很大差异,并且在储存和维护方面的复杂程度也不一样。而且,保持完整的、连续的项目参加者记录系统的可行性,也随干预的性质和可获资源不同而变化。例如,在医疗和精神卫生系统的案例中,复杂的、计算机化的管理和项目对象的信息系统因为照料的需要而得以发展,这是许多其他项目所不能够达到的。

专栏6—F 纽约和费城成年无家可归者公共避难场所的利用

费城和纽约市已将市政府资助或运作的庇护所的服务程序进行了标准化。所有获准进入公共庇护所的人必须为计算机化的注册提供进入信息,包括项目对象的姓名、种族、出生日期和性别,并且必须进行药物滥用和精神健康问题、医疗条件和残疾评定。从宾夕法尼亚大学来的研究人员进行的服务利用研究,分析了得自纽约市1987—1994年间(110 604名男子和26 053名女子)和费城1991—1994年间(12 843名男子和3 592名女子)的这些注册信息。

他们发现有三类主要的使用者:①长期无家可归者,特征是很少有不在庇护所的时候,时间可能持续几年;②间歇性无家可归者,特征具有多样性,从相当长的一段时间来看,呆在庇护所的时间越来越短;③过渡性无家可归者,他们往往在庇护所呆上一个或两个较短的时期。

最引人注意的发现是,长期无家可归者的数量和相对的资源消耗。例如,在纽约,18%的庇护所使用者在第一年里有180天或更长的时间呆在庇护所、消耗了项目为庇护所首次使用者准备的时间总数的53%,是他们应该呆在庇护所时间的3倍。这些长期使用者主要是老人和有精神健康问题、药物滥用者和一些有医疗问题的人士。

资料来源:Dennis P. Culhane and Randall Kuhn, "Patterns and Determinants of Public Shelter Utilization Among Homeless Adults in New York City and Philadelphia," *Journal of Policy Analysis and Management*, 1988,17(1):23-43. Copyright © 1998,John Wiley & Sons, Inc.

在测量目标人群参与的过程中,主要关心的是资料的准确性和可靠性。应当注意,所有记录系统都会有某种程度的错误。有的记录会包含不正确或过期的信息,还有的记录会不完整。不可靠的记录可被用于决策的程度取决于不可靠性的类型和程度以及决策的性质。很清楚,涉及重要结果的关键决策比不太重要的决策要求质量更好的记录。如果是有关项目是否继续的决策,就不应依据不那么可靠的记录,但是,这样的记录对于改变管理过程的决策,却已经足够了。

如果项目记录要在影响深远的决策中扮演重要角色,那么进行规则的记录核查通常是值得的。这样的核查与外聘会计师管理财会记录的目的是相似的。例如,可以通过抽样来了解,是否每个目标都有记录、记录是否完整以及是否遵守了记录规则。

调 查

使用项目记录评估目标人群的参与就是具体进行项目参与者调查。当所要求的资料不能通过项目日常活动来获得时,或者当目标群体的规模很大时,就应

该进行抽样调查,与全面调查相比,抽样调查更经济、更有效率。

例如,由父母发起的一个特殊辅导项目只在社区的少数学校开设。所有学校的孩子们都可以受到辅导,但项目员工没有时间或技能来管理适当的教育技能测试和完成给已经注册的孩子的记录工作。由于缺少完整的记录,评估小组就可以根据样本进行测试,估计所选程序的恰当性,并评估项目是否正在为目标人群服务。

如果项目不限于选定的、特定个体组成的群体,而是针对整个社区时,那么核查覆盖需求人群程度的有效方法,就是进行社区调查,这有时也是唯一的方法。各种健康、教育、娱乐和其他人类服务项目通常是社区范围的,尽管预期的目标人群可能有所选择,如过失青年、老人或育龄妇女。在这些案例中,调查是判断评估目标是否达成的主要方法。

对"感觉良好(Feeling Good)"电视节目的评估,例证了使用调查方法来为拥有全国观众的节目提供反馈信息的价值。这个项目是儿童电视工作室(the Children's Television workshop)(芝麻街的制片)的作品,用来激励成年人进行预防性健康实践。尽管对所有收入水平的家庭来说这个节目都能收到,但节目的主要目的还是激励低收入家庭的成年人加强他们的健康实践。盖洛普进行了4次全国调查,在"感觉良好"播映几周的不同时间里,每个时段都约有1 500名成年人在收看。这个资料提供了收看节目观众的规模和观众的人口、社会经济和态度特征资料(Mielke and Swinehart, 1976)。调查发现,项目根本没有到达目标群体,项目因此中止。

为了测量劳工部项目(如培训和公共部门雇用)的覆盖面,劳工部周期性进行全国抽样调查。收入和项目参与的调查由人口普查办公室执行,而社会项目的参与测量则由许多联邦部门进行。这些大型调查包括一个由21 000个家庭组成的、为期3年的追踪研究样本,通过个人访谈来判断每个样本家庭成年成员是否曾参加或正在参加诸多联邦项目的一种。通过对比项目参与者和非参与者,调查提供了项目覆盖面偏差的信息。另外,还产生了关于没有被覆盖、但有资格的目标人群的信息。

评估偏差:项目使用者、有资格者和中途退出者

项目参与偏差的评估可以通过检查参加项目的个体与要么中途退出、要么有资格而没有参加的个体之间的差异性来进行。简单地看,某个项目的中途退出率或人员缩减,可以作为项目对象对干预活动不满意的指标,也可表明社区条件对充分参与的影响。例如,缺少足够的交通就可能阻碍那些正在参与和有资格参与项目的人参与项目活动。

对完全不参与或不能充分参与项目的目标人群的特定次级群体进行说明,也是很重要的。这些信息不仅对判断努力是否值得有价值,而且对发展项目、修正假设(如何吸引和保留更多的目标人群)也是必需的。这样,参与的很多方面不仅对督导目标、而且对接下来的项目计划很重要。

中途退出者的资料可以来自服务记录或专门针对发现非参与者的调查。然而，社区调查通常是证明合格项目对象是否参与项目活动的唯一可行方法。当然，如果可以在评估项目绩效前获得整个合格目标人群的足够信息，就可以不进行这样的调查（譬如通过人口普查或筛选访谈获得资料）。无论搜集资料是出于项目规划的目的还是通过社区调查了解项目干预的状况，在获得资料之后，都涉及采用多种方法分析资料，从纯描述性的分析到很复杂的模型解释。

在第11章，我们将描述分析项目成本、收益的方法，这样的方法主要用来进行经济效率测量。很清楚，为计算成本，对需求人口或高危人群的规模、参与但中途退出的和一直参加到项目结束的群体进行估计，都是很重要的。同样的资料也可以用于估计收益。另外，在判断一个项目是否应该继续和是否应该在同一社区内或其他地方扩展时，这样的资料也很有用。而且，项目员工需要这样的信息，以完成他们的管理和解释责任。尽管项目参与的资料不能代替判断项目是否有效率、是否有绩效的信息，但有一点，如果没有这样的资料，也就无法正常分析项目的影响。

组织功能的督导

督导项目的主要组织功能和活动，侧重于项目是否在管理和使用资源以完成其基本任务方面运行良好。当然，在那些任务中，主要的是将预期的服务送达目标人群。另外，为了维持组织的可靠性和有效性，项目还必须实现各种支持功能，例如，基金筹集、督促、倡导、控制和管理。项目过程督导则探讨项目实际活动和安排是否完全符合预期。

在设计检测程序时，第5章描述的项目过程理论再次成为了有用工具。因此，组织计划就是相关部分的集成，一个清楚、充分表达的过程理论，将说明主要的项目功能、活动和产品，并且显示这些因素之间是如何彼此联系并与组织结构、员工模式和项目资源联系起来的。这个描述为评估者在说明重要项目功能及其发挥作用所需的先决条件时，提供指导。对于项目的有效执行而言，项目过程督导成为说明和测量项目活动和条件最关键的问题。

服务送达是基本要求

正如本章前面所提到的，对于许多不能显现影响的项目来说，问题就是没有送达项目设计中详细说明的干预，一般被认为这是执行失败。执行失败有三种：①没有或没有足够地送达干预；②送达了错误的干预；③干预没有达到标准或没有在控制范围内，并且在目标人群中有很大的差异。

“非项目的”和未完成的干预

首先考虑“非项目”问题（Rossi，1978）。麦克劳林（McLaughlin，1975）评论

了小学和初中教育法案第一号令的执行条款,这个法案每年拨出几十亿美元用于帮助地方学校解决贫困学生的教育剥夺问题。即使学校花费了这笔基金,地方教育权力机构也不能详细地描述他们的执行活动,并且很少能证明活动已送达给学校的目标群体。总之,没有证据能证明项目的存在合理性。

大量其他项目也有不能送达的情况。例如,达塔(Datta, 1977)介绍了职业教育项目的评估,并发现目标人群很少参加计划的项目活动。相似的情况是,针对一个用来激励有缺陷高中生向更高水平学习成绩发展的"提携"项目,评估结果显示,项目活动竟然主要由分发徽章和励志书籍构成(Murray, 1980)。

很大程度上不是完全没有送达服务,而是送达系统削弱了干预的强度,没有使足够数量的干预达到目标人群。问题可能在送达人员缺乏责任感,结果导致最小送达或把送达变成了仪式,此时项目实际上就已经不复存在了。例如,专栏6—G就是对第2章专栏的扩展,描述了福利改革的执行,在这个改革中,福利工作者与新政策的项目对象很少交流。

专栏6—G 第一线:福利工作者在执行政策改革吗?

在20世纪90年代早期,加利福利亚启动了工作报酬示范项目(Work Pays Demonstration Project),这个项目扩展了加州的职业培训项目(Job Preparation Program, JOBS),修正了家有小儿帮助项目(AFDC)的福利政策,以提高对寻求就业的刺激和支持。工作报酬示范项目被设计成是用来"基本改变AFDC项目的重点,促进工作而非福利、自给而非福利依赖"。

地方福利办公室的工作人员是工作报酬项目执行的关键点。他们的访谈量(包括受理和复查访谈)展现了福利系统与项目对象接触的唯一渠道。这个事实促使评估小组要研究福利工作者在其与项目对象互动期间是怎样传达工作报酬政策的。

通过"回推"法,评估者推断,福利工作者与项目对象涉及项目的交往包括某种"信息内容"和"正向判断"。信息内容是指向项目对象传递详细的项目信息;根据项目要求,福利工作者应该把工作报酬新项目的规则告知项目对象,说明把工作和福利结合起来所创造的自足机会更多,并把可靠的培训和支持服务告知他们。正向判断是指福利工作者有责任在知会项目对象时要采用教育、社会化的方式使项目对象明确与接受福利相关的预期和机会。项目期望福利工作者在项目对象访谈期间强调新的就业规则和收益,并让项目对象明白,福利只是他们在准备工作之前的暂时资助。

为评估福利工作者对新政策的执行,评估者观察和分析了工作报酬示范项目4个县的福利工作者与项目对象谈话中的66个,并制订了结构观察表,用于记录各种论题被讨论的频率和搜集个案的特征信息。这些观察的结果从两个维度编码:①信息内容,②正向判断。

结果表明,用评估者的话说:80%以上的谈话中,福利工作者没有提供和解释福利改革的信息。大部分福利工作者保持工具性处事模式,即强调福利工作者要搜集和证明项目对象合格性的信息。一些福利工作者则通过提供有关工作相关政策的信息来处理新需求而使信息常规化,并把其加入到标准化的、改编成可记诵的福利规则中来。还有的福利工作者通过将互动特殊化来处理,在此基础上,给部分项目对象一些当时的信息。

这些发现意味着,福利改革在加州这些县的街道没有得到充分执行。福利工作者与项目对象的关系模式包括福利申请过程、合格规则的执行和诸如JOBS服务等稀缺资源的配置,与新的强调外在于福利系统的工具性资助、工作和自足不太一致(第18-19页)。

资料来源:Marcia K. Meyers, Bonnie Glaser, and Karin MacDonald, "On the Front Lines of Welfare Delivery: Are Workers Implementing Policy Reforms?" *Journal of Policy Analysis and Management*, 1998, 17(1):1-22.

错误干预

第二类项目失败(即错误干预的送达)有几种情形:一种是送达方法否定了干预。一个例子就是进行合同实验,在这个实验中,根据合同讲授数学和阅读的私人公司,按学生取得的成绩被支付报酬。但这些公司在学校送达这项服务时有很大的困难。在一些地方,学校系统破坏实验;另一些情况是,公司面临设备不足和教师不友好的困境(Gramlich and Koshel, 1975)。

另一情形是,所要求的送达系统太复杂,进而导致了错误干预。这样的情形在探索性项目和全方位的复杂项目执行之间是有差异的。当要求员工具有足够的技能训练和士气才能运作的巨大送达系统被交给技能训练不够和士气低下的员工进行管理时,就只能以失败而告终。教育领域再一次提供了一个例子:教育方法,如计算机辅助学习或个别辅导,在实验发展中心,这样的活动运行良好,但在学校系统则不灵了。

干预与送达方式之间的区别并不总是清晰的。这种差别在收入维持项目中相当清楚,在此,"干预"就是给受益人金钱,送达模式则包括直接存入对象户头或者把支票、现金送到对象手里。项目想要的把钱送到对象的手里;如果只是转账(无论是电子还是手工),就没有干预的效果。相反,咨询项目就可以通过重新培训目前的员工、雇用咨询者或雇用合格的心理治疗专家来处理;在这种情况下,治疗和送达方式之间的差异是模糊的,因为咨询方法随咨询者的不同而不同。

不合格的干预

最后一类执行失败是不合标准的或没有得到控制的干预。如果项目设计中有关送达的部分留下了太多的随意性,就会在不同的地方产生重大差异。经济机会办公室(the Office of Economic Opportunity)的早期项目提供了这方面的例子。社区行动项目(the Community Action Program, CAP)给地方社区在选择各种行动时留下了相当大的自主性,而只要求在贫困地区"最切实可行的参与"。由于不同城市的项目结果不一致,要把 CAP 项目所完成的事项编成文件几乎不可能(Vanecko and Jacobs, 1970)。

同样,早期教育项目(project head start)给地方社区一定资金,让他们为贫困家庭儿童进行学前教育。在最初的十年内,覆盖面、内容、员工资格、目标和一些其他特征,每个学前教育中心都不相同(Cicirelli, Cooper, and Granger, 1969)。因为没有具体的早期教育设计,所以无法对项目样本是否获得了成功做出结论,所能做的只能是一般化的结论,即一些项目有效果、而另一些则没有,在有效果的项目中,一些比另一些更成功。直到最近十年,才实现了一定程度的标准化;

再到2001年时,设计和开展项目绩效评估才得以可能(Advisory Committee on Head Start Research and Evaluation,1999)。

送达系统

项目送达系统是提供干预时所采取的方法和行动的结合,通常由许多独立的功能和关系模块构成。作为一般规则,明智的做法是评估所有的因素,除非以前送达系统某方面的经验表明,对这个方面的评估没有必要。两个概念对督导项目送达系统的执行特别有用:服务的专门性和可及性。

服务的专门性

对计划和督导目的而言,将所提供的实际服务具体化是很有意义的。这一工作需要将项目实际提供的服务以操作化的(可测量的)方式细分。第一个任务是根据所从事的活动和参与的提供者来界定每一种服务。如果可能,最好把项目的不同方面区分为独立的、独特的服务。例如,如果一个为学校中途退学的学生提供技术教育的项目包括文化培训、木工技能和在职学徒工作,那么就要将这三个方面分开督导。此外,为了估计项目成本(便于进行成本收益分析和财务记账),常常要给不同的服务进行货币性赋值。如果要比较几个项目或者根据提供的服务量报账,这一步就很重要。

对于项目督导来说,简单具体的服务比较容易说明、计算和记录。然而,如果要设计与项目目标一致的执行,通常要涉及复杂的因素。例如,儿童门诊要求准许参加体检,但体检的范围和内容却取决于每个儿童的特征。这样,"体检"就是一种服务,如果不对每个接受检查的儿童界定不同的服务,那么活动的组成部分就不明了。战略性的问题就是如何打破均衡、界定服务,以使独特的活动能被可靠地说明和计算。同时,就项目的目标来说,这种区分也是有意义的。

在干预性质要求项目行动范围广泛的情况下,就只能根据服务提供者的一般特征和花在服务活动方面的时间来描述服务。例如,如果一个项目让低收入社区的熟练女工匠来教导社区成员,以期改善他们的居住环境,那么这位女工匠的具体活动就有可能在不同的家庭之间产生巨大差异。她们可能建议某家庭如何设计窗户,而建议另一个家庭如何加固房基。试图把这些活动编成文件的督导计划,只能用一般性的术语和用例子来描述这些活动。然而,还是有可能把提供者的特征和她们花在每种服务对象上的时间具体化,例如,在房屋建筑和维修方面有5年的经验,具有木工、电线配置、基建和外围建筑方面的知识。

实际上,服务通常根据时间、成本、程序或产品的单位来界定。在职业培训项目中,服务计量单位是指提供咨询的小时量;在住房改进项目中,则根据提供的建材数量来界定;在家庭工业项目中,服务单位是指活动,比如如何操作缝纫机的训练课程;在教育项目中,单位可以是具体课堂里具体课程材料的使用情况。所有这些例子要求对服务构成有明确界定,对具体服务而言,单位就是能恰当地描述服务数量的范畴。

可及性

可及性(Accessibility)指结构的和组织的设置有利于提高项目参与的程度。所有项目都为把服务提供给恰当的目标人群而准备了一些策略。在一些例子中,可及性仅意味着设立办公室和在这样的假设下运行:预定的参与者会"自然地"来利用办公室所提供的服务。然而,在另一些例子中,为保证可及性,则要求展开招集参与者的运动,把参与者带到干预地点,并在干预期间最大限度地减少中途退出者。例如,在许多大城市里,会在寒夜派特殊小组到街道说服睡在露天的无家可归者到庇护所过夜。

在涉及可及性时,大量的评估问题便出现了。一些仅与服务送达相关,一些与以前所讨论的服务利用类似。主要问题是,项目行动是否与项目设计和预期一致。例如,在居住着大量西班牙人的地方,精神健康中心是否总是有说西班牙语的员工?

潜在目标人群也匹配了相应的服务吗? 例如,据观察,最初利用紧急医疗照料的社区成员,可能在以后将其当作了一般性的医疗照料。这种紧急服务的成本很高,如果误用,就会降低其他社区成员的利用的机会。一个相关的问题是,根据特定社会、文化和种族群体的目标,可及性战略是否鼓励差异对待,或对所有潜在目标人群都平等相待。

项目支持功能

尽管提供预期服务被假设为项目的主要功能,而对于督导者来说,重要的一点是,大部分项目也具有重要的支持功能,这些功能对维护自身的能力和继续提供服务来说是关键的。当然,项目管理者对这些功能很有兴趣,但也通常与评估者或外部决策者的督导有关。重要的支持功能包括诸如基金筹措这样的活动;公共关系可以提高潜在资助者、决策者或一般公众心目中的项目形象;还有,员工培训包括直接服务员工的培训、招集和保留关键员工;发展和维护与相关项目、资源等的关系;获得服务所要求的物资;倡导代表目标人群的利益。

项目督导计划能够而且通常应该把重要的项目支持功能指标与服务活动相关指标结合起来。在形式上,这些指标以及识别这些指标的过程与描述项目服务没有太大的不同。首先必须识别关键的项目活动,并把项目活动具体、清楚地表述出来,譬如基金筹措活动、已筹到的资金、培训课程、支持性事件,等等。接下来,就要使测量能够区分好的与差的执行,并能有规律地进行信息搜集。这些测量被包括在项目督导程序和涉及项目绩效的其他程序中。

项目过程督导资料的分析

当然,只有恰当地分析资料,资料才会有用。一般而言,督导资料的分析可

以说明以下三个问题:项目绩效描述、场所之间的比较和项目与设计的一致性。

项目绩效的描述

要评估项目按设计执行的程度,依赖于对项目实际运行的充分和准确描述。一个源自督导资料的描述应包括以下主题:覆盖面的估计和参与偏差、送达服务的类型、给主要参与者的服务强度和参与者对所送达服务的反应。描述性陈述可以采取叙述性的形式,尤其是当督导资料来自定性来源或者以表格、图表等形式所概括的定量来源时。

场所之间的比较

当一个项目包括一个以上场所时,第二个问题就是关心项目绩效在场所之间的差异。场所比较提供了对项目实施差异和最终产出差异多种来源的信息,例如,员工、管理、目标人群或周围环境的差别,同时,这样的比较也有利于促使项目达到标准。

项目执行与设计的一致性

第三个问题就是我们开始提到的:项目的设计与执行的一致程度。两者之间产生差距是不可避免的,如果项目功能或者运行状况不像所期望的那样好,都可能会造成差距。如果有这样的情形,人们就可能努力按照起初的设计执行项目或进行重新设计。这样的分析也为判断影响评估的恰当性提供了机会,如果必要的话,也为用更规范的评估来评价设计和执行提供了极好机会。

小　结

- 项目督导是评估的一种形式,用来描述项目是如何运作的、评估项目实现预期功能的程度。项目督导建立在项目理论基础上,而项目理论则用于识别假定的、使项目有效果的、必要的重要组成部分、功能和关系。
- 有了项目督导结果,就可以针对项目理论、管理标准、应用法规、伦理、专业标准和事后判断的要求进行绩效评估。
- 项目督导的一般形式包括过程(或执行)评估、管理信息系统和绩效测量。
- 过程评估用于评估服务是否按预期送达项目对象,这样的评估常常由评估专家独立进行。如果问题仅仅是项目运行、服务送达和其他类似问题,那么就可以是独立的评估。过程评估也通常伴随影响评估,以确定项目提供了什么服务和服务具有什么影响。
- 如果项目有着完善的信息管理系统(MIS),那么就能够把项目督导整合进项目日常信息的搜集和报告之中。
- 如果从评估、责任和管理的立场进行项目督导,则有不同的形式和不同的目标,但所需资料的类型和搜集资料的程序和方法则相同地或相当程度地重叠。尤其是,一般都包括项目绩效中的一个或两个领域:服务利用、组织功能。

- 服务利用问题一般包括覆盖面和偏差。涉及覆盖面的有用资料主要是项目记录、项目参与者调查和社区调查。偏差则能通过项目使用者、合格的非参与者和中途退出者的比较来揭示。
- 对项目组织功能的督导所关注的是项目如何组织其努力和使用其资源来完成基本任务。特别要关注的是识别在项目执行中阻止为给目标人群送达服务的缺陷。执行失败的三个来源是:不完整干预、送达错误干预和不符合标准或没有控制的干预。
- 对组织功能的督导也关注送达系统和项目支持功能。
- 督导资料的分析一般强调这样一些问题:项目运行的描述、场所比较、与原设计相符性和与标准和期望相比较的项目绩效。

基本概念

可及性(Accessibility):项目组织性和结构性安排的服务的可用程度。

责任承担力(Accountability):项目人员的责任就是为项目各方和主办方提供资料,以说明项目是有效的和与其覆盖面、服务、合法性和财务的要求相一致的。

管理标准(Administrative standard):由项目管理者或其他责任主体所设置的衡量尺度,例如,一个月内要接收90%的服务对象。这些标准设置的依据是过去的经验、可比项目的执行情况或专家的判断。

偏差(Bias):目标人群中次级群体接受项目干预的覆盖面差异程度。

覆盖面(Coverage):项目达到预期目标人群的程度。

管理信息系统(Management information system, MIS):通常是计量化的资料系统,例行搜集和报告关于服务送达给项目对象的信息,通常包括支出表、成本、诊断和人口特征信息以及产出状态。

产出督导(Outcome monitoring):测评并报告项目要提高的社会条件的状态的指标。

项目过程督导(Program process monitoring):长时期、反复地对项目过程进行评估。

7 项目产出的测量和督导

前一章讨论了如何督导项目的过程和绩效。然而，所有项目的最终目标并非仅仅为了运行良好，而是带来变化——以有益的方式影响某些问题或者社会状况。改变状况是项目的预期产出。评估项目获得这些产出的程度是评估者的主要职责。

在项目影响理论中经常可以看到项目预期产出。对产出进行敏感而有效的测量不但在技术上具有挑战性，更重要的是它是项目评估获得成功的关键。另外，持续的产出督导对于项目的有效管理也十分重要。然而，解释项目产出的测量和督导的成果，对于项目方来讲却是一个巨大的挑战。因为，某些项目产出可以由其他因素而不是项目过程所产生。本章描述了如何识别项目产出，怎样测量和督导项目产出，以及如何恰当地解释这些结果。

评估具体干预项目对项目对象和其致力改善社会状况的效果，是最重要的评估任务，是项目评估合法性的保证，因为它涉及社会项目的“底线”。不论具体项目在描述目标群体需求、制订完善的实施计划、介入目标人群和送达适当的服务上做得如何好，都不能据此判定项目的成功，除非项目活动的确带来了既定社会领域的有益变化。因此，对有益变化的测量不仅仅是评估的主要功能，也是评估项目中高难度、高风险的活动。鉴于这些原因，很显然，评估者必须承担的职责是：小心谨慎地完成这些任务，确保评估结果的有效和相应解释的适当。由于这些原因，这也是评估者承担的最困难任务之一，往往带有“政治任务”的意味。

在本章和第10章中，我们考虑怎样最好地鉴别项目所预期产生的变化，怎样设计这些变化的测量标准，以及怎样解释测量标准。讨论项目的影响，首先要理解项目的产出，所以我们首先讨论这个关键的概念。

项目产出

产出，是由于社会项目的干预所导致的目标群体状态或目标社会状况的改善。例如，中学反吸烟运动爆发后，青少年吸烟人数就是一个产出。那些尚未开始抽烟的人对待抽烟的态度也是一种产出。同样，参加过学前教育项目的儿童的“学前准备”也将是一个产出。还有，完成减肥计划者的体重，参加过管理培训项目的职员的管理技巧，以及经过地方环境保护组织打击污染行为后当地河水中污染物质的数量，都是产出。

请注意这些例子的两个方面。首先，产出是目标人群的或社会状况的而不是项目的可观察特征。项目产出的界定和项目行为没有直接的关系。尽管直接指向项目参与者的服务经常被描述为项目的“输出”，也就是这里所说的产出，但项目产出必须与项目能够为参与者提供的产品或服务获益相关，而不仅仅是“收到”产品或服务。因此，“接受家庭支持治疗”不是我们所说的项目产出，而是项目服务的送达。同样，向100户不便于外出老年人提供膳食不是一个项目产出，而是服务送达，是项目过程的一个方面，膳食对于老年人健康的营养价值才是产出。还有民心的提升、可观测的生活质量、亲自下厨的受伤风险等都是产出。此外，项目产出原则上通常可以通过反观未接受项目服务的个人未受干预的状况而比较出相关特征。例如，我们可以在没有项目干预的情境下评估抽烟的数量、学前准备、体重、管理技巧以及水污染，然后，在项目实施一段时间后，进行一次新的测量。的确，正如我们稍后会谈到的，我们可以将没有项目干预下的测量结果和项目执行中得到的测量结果进行比较。

第二，正如我们界定的，项目产出概念并不必然意味着项目的目标对象已经实际发生变化或者项目已经导致它们以某些方式发生变化。禁烟运动爆发后，中学生抽烟的数量可能没有马上开始变化，也可能并没有人在加入减肥项目之后体重立刻减轻。除此之外，还可能发生项目预期方向之外的变化——青少年

抽烟人数可能增加,项目参与者可能增重。如果我们深究其因,就可以了解到,这些变化可能是其他因素而非项目影响所致。或许,减肥项目是在人们易于过度摄入甜食的假期进行的;而青少年抽烟人数的减少可能是对摇滚音乐名星因抽烟死亡的事件性反应。因此,评估者面临的一个挑战就是:不仅要评估项目实际获得的产出,而且还要评估产出中何种程度的变化是由项目活动所致。

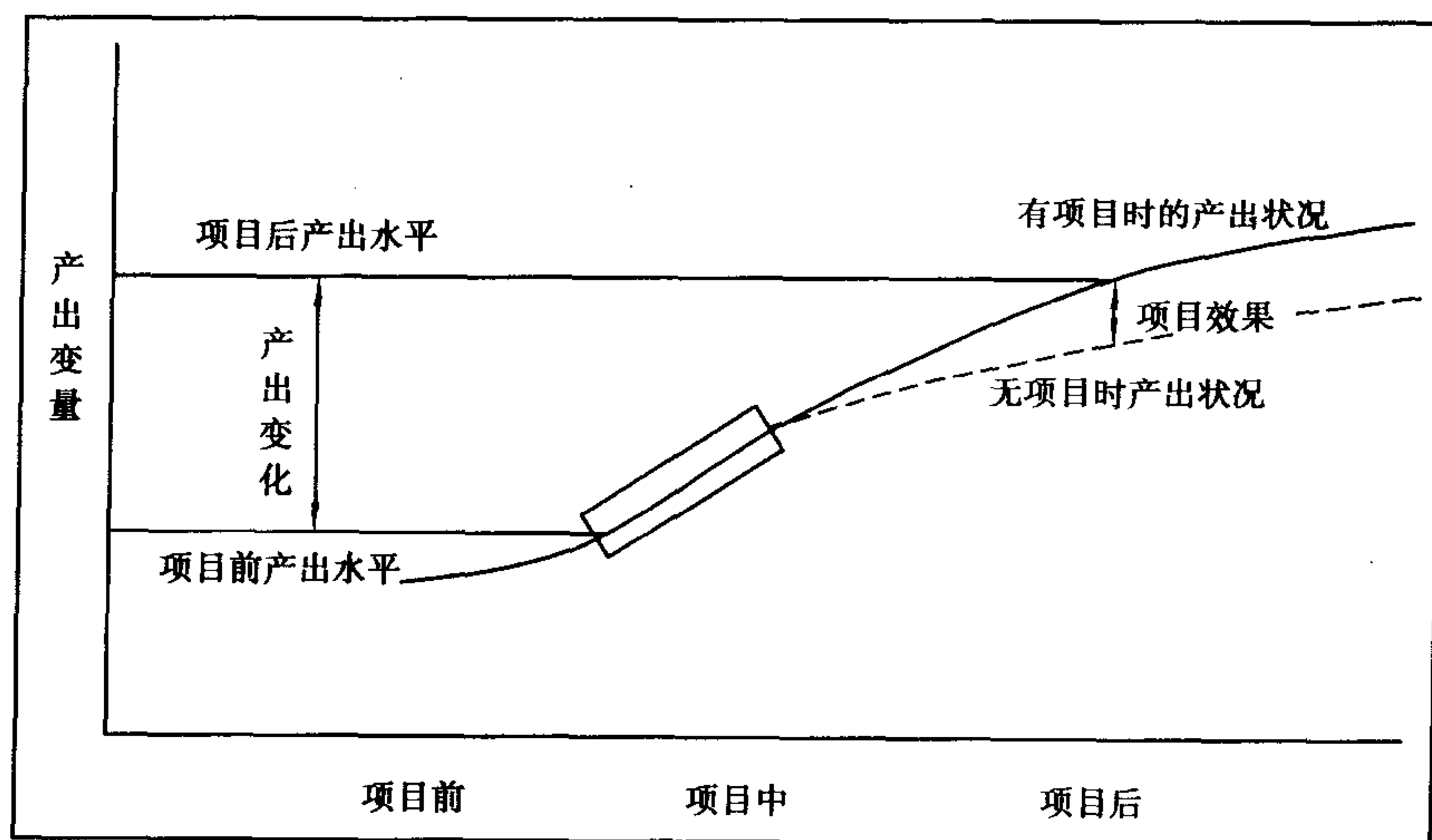
产出水平、产出变化和净效果

在下面的讨论中,我们要特别注意“产出”术语的重要区分:

- **产出水平**(Outcome level)是指某一时点的产出状况。
- **产出变化**(Outcome change)是指不同时点的产出水平差异。
- **项目效果**(Programe effect)是因项目活动导致产出变化的部分,与其他因素引起的变化相对应。

专栏 7—A 的曲线绘出了产出测量水平的历时变化。纵轴表示我们希望评估的、与项目相关的产出变量。产出变量是项目目标人群的可测量特征或条件,这些目标人群可能受项目活动影响。产出变量可能是抽烟人数、体重、学前准备、水污染程度,或任何其他我们上面所界定的产出。横轴表示特定的时间,从项目开始之前到项目执行之后的一段时间。实线显示接受项目服务个体的平均产出水平。值得注意的是,他们的历时性状况并不是水平直线,而是一条摆动的曲线。这暗示着抽烟、学前准备、管理技巧以及其他此类产出维度不像预期那样保持不变——他们的变化是由诸多来自项目之外的自然因素和环境造成的。例如,从前青春期到青春期的几年内,抽烟人数会随着年龄增长而增加。水污染水平可能会因当地人的行为和气候条件而上下波动,例如大雨可以稀释污染浓度。

专栏 7—A 产出水平 产出变化和项目效果



如果我们测量产出变量,就可以判断目标人群测量值的高低,例如抽烟人数

或学前准备就绪人数。这将告诉我们产出水平，经常被简称为产出。当对接受项目服务的目标人群进行测量时，我们可以得到相关的信息——多少青少年抽烟、学前儿童中“学前准备”的平均水平、水里有多少污染物质。如果所有青少年都抽烟，我们可能会很失望；相反，如果青少年中没有人抽烟，我们可能会很高兴。尽管可以限制事情发生的可能性，但是，这些产出水平本身并不会告诉我们项目的有效程度。例如，如果所有十多岁的孩子都抽烟，我们可以相当肯定地认为反抽烟项目并不成功，或者可能具有反作用。如果十几岁的青少年都不抽烟，这就强烈地暗示项目已经发生作用，因为按照常规，不会所有人都自愿放弃抽烟。当然，这类极端的结果相当少见。因此，在大多数情况下，单有项目产出水平，并不能解释项目的成败。

如果我们可以测量目标人群在参加项目前和参加项目后的结果，那么可以描述的内容不会限于产出水平，我们还可以识别出产出变化。假设专栏7—A中的图表展示给我们的是一个学前项目中的儿童在上学前的准备状况，显然，儿童在参加项目之前具有较少的准备，参加项目之后则有较多的准备，这就是项目带来的积极变化。即便他们在项目之后的学前准备状况并不如学前老师预期的高，但是已有变化已经显示了在学前准备方面的积极变化。当然，仅从这一信息，我们并不能确定前项目与儿童学前准备提高之间有联系。学龄前儿童处于快速发展阶段，他们的认知和动机本来就可以通过正常的成熟过程迅速增长。此外，其他因素也可能发生作用，如他们的父母可能为他们阅读，帮助他们开发智力并为入学做准备，这些也是孩子们准备状况提高的部分原因。

专栏7—A中的虚线表示产出变量的变化轨道，如果项目参与者没有接受项目，其变化就是这条虚线。例如，对于学前儿童，虚线表示没有参加学前项目的儿童上学前准备状况的变化。实线表示参加项目的儿童上学前准备状况的发展。两条曲线的比较显示出，即便没有参加项目，儿童上学前准备状况也有提高，虽然不如参加项目儿童提高得显著。

同一个人参与项目与不参与项目所得到的产出测量值之间的差异，是项目产生的产出变化。这是在项目之外无法发生的项目干预的附加价值或净收益。我们将这个增量称作项目效果或项目影响。不过，它仅是测量产出的一部分，也是项目得到好评的前提条件。

项目效果评价，或影响评估，是最苛刻、最严格的评估研究任务。实际上，专栏7—A已经凸现出了这种困难，因为项目效果不同于实际测量到的产出，测到的结果可以在项目之外产生。很显然，我们不可能观察到同一批人（或其他整体）同时处于参与和不参与项目的产出状态。因此，我们必须观察参与项目后的产出，然后再评估没有参与项目可能产生的相应产出。由于对没有参与项目可能产生的相应产出进行了假设，实际上项目对象可能接受了项目干预，所以产出测量值是推断值，而非真实的测量值或观察值。在这些条件下，进行合理推论是困难的和成本高昂的。第8章和第9章描述了评估者针对这项挑战性任务所采用的方法和工具。

尽管产出水平和产出变化对确定项目效果的使用价值十分有限，但对管理者和项目主办方督导项目实施却具有重要价值，我们会在本章稍后的内容中讨

论这一应用价值。现在,我们继续探讨产出概念,通过讨论项目产出的识别、界定和测量来实现评估目的。

识别项目产出

进行项目产出测量的第一步,是非常具体地界定要测量的产出变量。为此,评估者必须考虑项目方对项目预期产出的看法,这些预期产出在项目影响理论和相关研究中已有详细说明。另外,评估者也需要关注项目实施可能产生的意外产出。

项目方的视角

不同的项目方对于"项目应该完成什么"有自己的理解,相应地,他们在期望项目将产生什么样的产出方面也有不少差异。项目预期产出的直接信息,通常是项目目的、目标和任务。来自于项目主办方的服务协议和投资计划,常常是项目所期望的结果。

这些信息通常有一个缺陷:没有清晰地界定具体产出测量的特异性。因此,评估者首先要将项目方的原始信息转化为可操作的数据形式,并和相关项目方商榷,以保证产出测量符合他们的期望。

出于评估者的目的,产出描述必须显示项目所期望变化的特征、行为或状况。然而,如前所述,从产出描述到选择或实施产出测量,评估者必须进一步地说明和区分项目的产出。专栏 7—B 展示了一个常用的产出描述例子。

专栏 7—B 描述具体的、可测量的几个项目产出例子

青少年犯罪:根据法律可以提起起诉的、由 18 岁以下青少年实施的犯罪行为,无论是被执法机关查出的还是青少年自首的。

与反社会伙伴接触:朋友式的互动,与一个或多个同龄少年一起经常从事的不法行为和/或对他人有害的行为。

积极地利用休闲时间:在学校和工作之余时间内,从事具有教育意义的、社会性的或具有个人价值的活动。

水质:由于水中存在某些物质,对人类和其他饮用此水或与此水有接触的现存生物体十分有害。

有毒物质泄漏:工业设施中对环境有害的物质排放,在某种意义上,指将人类和其他现存生物暴露于这些有害物质面前。

认知能力:在思考、问题解决、信息处理、语言、精神印象、记忆和整体智力等方面的表现。

入学前准备:儿童入学时的学习能力,尤其是身体健康和身体发育、社交和情感的发展、语言和交际技巧、认知技能,以及使儿童能够在正式的学校教育参与中受益的一般性知识。

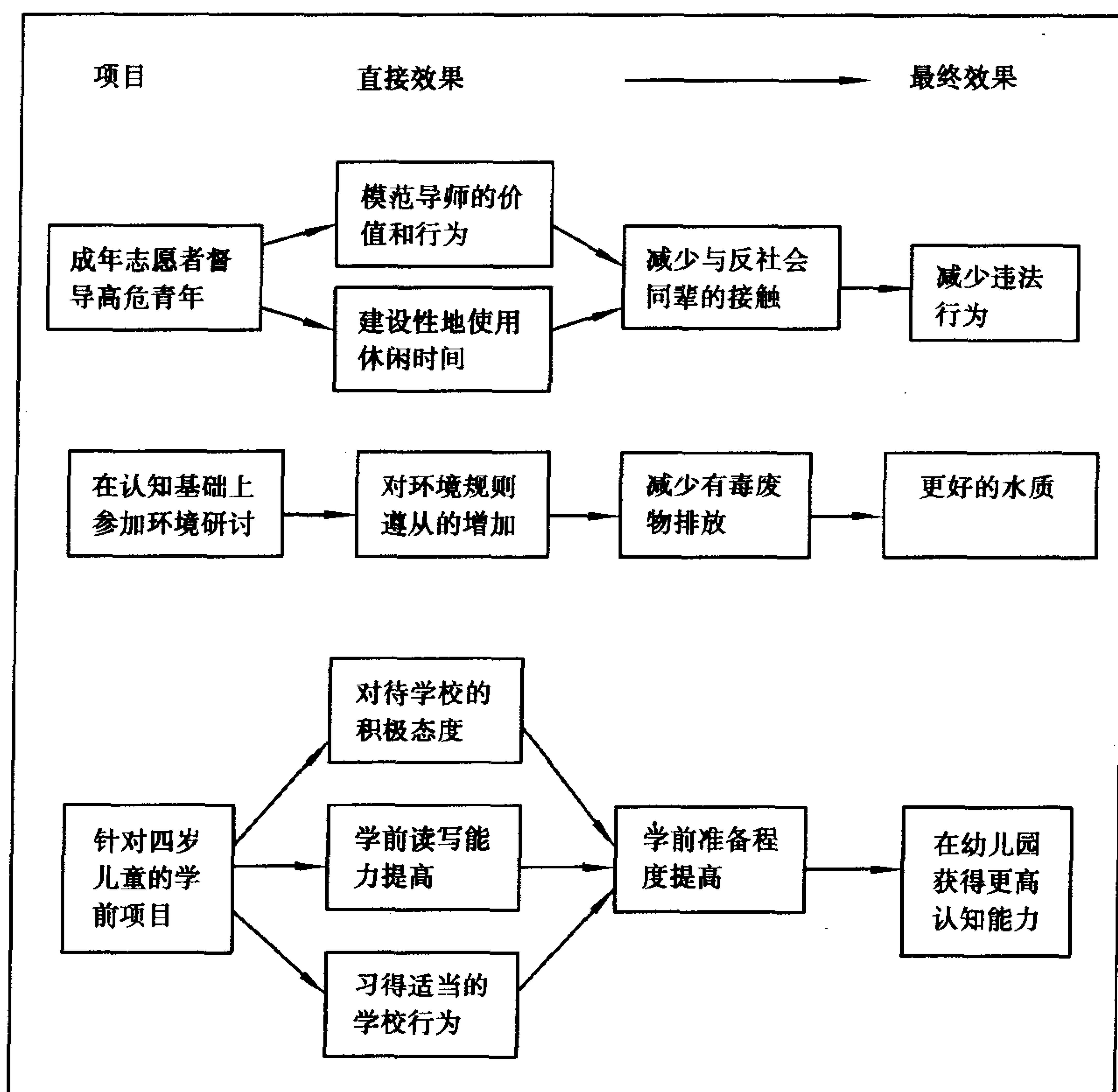
对学校的积极态度:儿童对学校的喜爱,对学校活动参与的正面感觉,以及参加学校活动的意愿。

项目影响理论

正如第5章所述，全面而清晰的项目影响理论，对于界定和组织项目产出尤其有用。影响理论将社会项目产出表述为项目逻辑结构的一部分，即将项目活动与最直接的项目产出连接起来，顺其自然，人们期望由此导致更加长远的影响。如果表述正确，产出间的一系列相关关系代表了项目假设，各种假设呈现了项目服务与项目可能产生的社会收益之间的关联。因此，当界定应纳入测量的产出时，对评估者总结项目理论而言，项目假设尤为重要。

专栏7—C展示了几个影响理论的项目逻辑模型部分的例子（其他的例子在第5章中可以找到）。为了评估项目产出，在这些产出次序中，界定直接的和最终的产出之间的不同特征是有益的。直接的产出是指最直接和最迅速产生影响的项目服务，即所谓“取来”的产出，这些产出是项目参与者参与的直接成效，当他们离开时产出也会随即消失。对于大多数社会项目来说，这些最近的产出是心理的——态度、知识、意识、技巧、动机、行为意图，以及其他易受项目过程和服务直接影响的东西。

专栏7—C 几个显示直接的和最终的项目效果项目影响理论



通常,直接产出很少是项目的最终结果,正如专栏7—C中的例子所展示的那样。从社会或政策的视角来看,仅此而言,它们并不是最重要的项目产出。然而,这并不意味在评估中应该被忽视。这些产出是项目最可能获得的,所以了解项目是否获得这些产出上的变化,是很有价值的。如果项目没有产生这些最迅速和最直接的产出,并且项目理论是正确的,那么逻辑链条中的长期产出就不可能发生。另外,直接产出最易测量,也最容易归结为项目绩效。如果项目的这些产出是成功的,就可以确证项目活动是可信的。比较难于测量和归因的是较为末端的产出,对项目的长期效果,一般不太明确,也不易界定。如果有直接产出的信息,那么,长期成效就会更加清晰和易于评判。

但是,最终产出显然才是最具实践和政策重要性的项目效果。因此,明确地界定和描述项目活动预期产生的这类产出尤为重要。出于这些目的,认真建构项目影响理论,在项目性质一定的前提下,无疑将为评估产出的合理性提供基础。

然而,在影响理论中,项目通常会对最终产出产生更小的直接影响。另外,最终产出也会受到项目之外诸多因素的影响。这一状况使得界定预期的最终产出显得十分重要,即以某种方式将最终产出与项目活动可能影响的社会状况的各个方面紧密关联。例如,一个针对小学生的辅导项目可能主要关注阅读问题,意在提高这方面的教育效果。要评估这一项目的产出,就应该区分与读书技能紧密相联的领域和基本不受项目影响的领域,例如数学。

已有研究

要识别和定义项目产出,评估者就应该彻底核查与被评估项目相关的已有研究发现,尤其是回顾对相似项目的已有评估研究。了解哪些产出在其他研究中已经得到检验,同时,注意那些以前可能被忽视的相关产出。判断已有研究如何界定和测量产出,也是十分有用的。在某些评估领域,已经有着相对标准化的界定和测量工具,而且具有确定的政策意义。在另一些情况下,评估者则要了解某些界定或测量方面所存在的、大家都知道的问题。

意外产出

到目前为止,我们已经考虑了如何识别和界定项目方期望项目产生的产出和项目影响理论中的逻辑产出。然而,项目也可能获得明显的意外产出,而且上述方法无法识别这些意外产出。意外产出可能是积极的,也可能是消极的,其独特之处在于它产生于项目设计和直接意图之外。正是由于这一特征,使得意外产出非常难于预见。相应地,评估者必须花费专门的精力来识别任何潜在的意外产出,这对于评估社会项目的影响具有重要意义。

已有研究对于某一问题的分析通常很有帮助。也许,在类似项目环境下,已经有一些研究者就可能产生的净项目效果进行过讨论。在这一方面,不仅相关的其他评估研究,而且任何与手头项目要干预的社会状况相关的研究也被包括

在内。比如,对于毒品使用和使用者生活状态的研究,同样可以为某一项目干预提供线索,而这些线索很可能是项目规划所未考虑到的。

通常,某些人可以观察到意外的产出,从他们那里常常可以获得有益信息,他们的意见和说明很有价值。鉴于这一点,正如在本章提到的,与所有级别的项目人员、项目参与者和其他对项目有看法的关键人进行实质性的接触和沟通,对于评估者十分重要。如果意外产出在项目运作中相应而生,那么评估系统中一定有人意识到并且可以提醒评估者留心。这些人可能不使用"意外产出"的专业语言,但是,他们对项目相关方面的见闻和经历的描述,可以帮助评估者意识到意外产出可能是项目逻辑或主要项目方意图之外的重要效果,评估者可以根据情景对这些意外产出进行解释。

测量项目产出

并非所有已经描述的在项目过程中产生的产出都具有同等的重要性或适当性,所以评估者不需要通过测量所有产出来得出对项目干预效果的评价。因此,有必要进行某些选择,即只对其中一部分产出进行分析。另外,一些重要的产出,例如,长期产出——测量起来可能十分困难或成本偏高,通常也不包括在评估活动中。

一旦选定相关的产出,并已掌握全面和细致的描述,评估者必然面对如何测量的问题。产出测量是通过在项目环境中不断变动和系统变化的可观测指标来呈现项目产出的状况。一些项目产出与简单的容易观察到的境况相关,实际上项目产出测量经常是单维度的。例如,一个工业安全项目的预期产出可能是工人在工作场所带上护目镜。评估者在任何时间,通过简单观察和记录工人是否戴护目镜,并且通过定期的观察、扩展观察以识别工人带护目镜的频率,从而很好地描述项目产出。

然而,很多重要的项目产出并不像工人是否戴护目镜这么简单,是多维度的。为了完整地表述项目产出,将产出视为项目试图产生的多维度效果是十分必要的。例如,专栏7—B就提供了一个从法律控制犯罪视角,对犯罪青少年的描述。然而,青少年的违法犯罪,在几个维度上会受到"减少犯罪"项目的影响。首先,犯罪的频率和犯罪的严重程度之间是互相联系的,犯罪频率降低而严重程度上升是项目人员不乐意看到的。其次,也需要考虑犯罪的类型。例如,关注毒品滥用的项目可能认为毒品犯罪是最重要的相关产出,核查财产犯罪也是很明智的一个维度,因为毒品滥用者通常以财产犯罪来支持其毒品消费。其他犯罪类型或许也是相关的,但相关程度要小得多,这样就可把所有其他犯罪类型混合在一起作为单一的产出进行测量,而不需要在其间进行区分。

大多数项目的产出测量是多维度的,所以评估者需要考虑项目效果的多个方面或诸多成分。一般来说,评估者应该尽可能地确保不忽略项目产出的所有

重要维度。但这并不意味着所有维度都必须得到同等的关注,或者把所有维度都纳入产出测量范围。这一问题的重点是,评估者在确定最终的测量标准之前,应该考虑所有潜在的相关维度,并进行重要性的区分。专栏7—D列举了几个需要考虑多方面和多个维度项目产出的例子。

项目产出多维度的涵义之一是指:单一的产出测量可能并不能充分地代表项目所有的特征。例如,在青少年犯罪的例子中,评估者可能同时使用犯罪频率、犯罪严重性、干预之后第一次犯罪的时间和犯罪类型作为产出测量变量,用来全面评估项目干预的产出。的确,项目重要产出维度上的多元测量可以帮助评估者减少失误和偏差,因为有限的测量工具和标准总会遗漏掉某些与项目相关的重要产出维度。

专栏7—D　构成项目产出多元维度的几个例子

青少年犯罪

- 一段时期内可被指控的犯罪数量
- 犯罪严重性
- 犯罪类型:暴力犯罪、财产犯罪、毒品犯罪、其他类型犯罪
- 官方对犯罪的反应:警察接触或逮捕;法庭判决、定罪或者处置

有毒废物排放

- 废物类型:化学的、生物的;特殊毒素
- 废弃物质的毒性和危害性
- 一段时期内流出废物的数量
- 废物流出的频率
- 废物流出与居民区的接近程度
- 毒素通过蓄水层、大气、食物链和此类环节的散布速度

对学校的积极态度

- 对老师的喜爱
- 对同学的喜爱
- 对学校活动的喜爱
- 上学的意愿
- 自愿参加学校活动

同时,多维度测量也可以降低效果不佳的测量所带来的描述不足。如果不测量项目最可能施加影响的产出,会使项目活动看起来并不像实际那样有效。产出测量依赖于观察,例如,使用多个观察者可以避免个人偏见。如果评估儿童与同伴的侵略性行为就需要家长的观察、老师的观察,以及任何处于某一位置并可以发现儿童行为重要内容的人们的观察。专栏7—E将具体讨论多维度测量的一个例子。

专栏7—1 产出的多元测量

俄勒冈州阻止青少年使用烟草的社区干预,包括青年人反对烟草活动(例如,海报和免费体恤)和家庭沟通活动(例如,给家长散发宣传手册)。在影响评估中,从多个维度测量了项目产出。

青年的产出

- 对于烟草使用的态度
- 关于烟草的知识
- 关于烟草与家长的评论
- 抽烟或食用烟草的既定态度
- 上个月是否抽烟或食用烟草,如果是,有多少

父母的产出

- 关于烟草的知识
- 对待社区干预烟草使用的态度
- 对于烟草使用的态度
- 与儿童讨论不使用烟草的意愿
- 与他们的孩子讨论不使用烟草的谈话频次

资料来源:A. Biglan, D. Ary, H. Yudelson, T. E. Duncan, D. Hood, L. James, V. Koehn, Z. Wright, C. Black, D. Levings, S. Smith, and E. Gaiser,"Experimental Evaluation of a Modular Approach to Mobilizing Antitobacco Influences of Peers and Parents",*American Journal of Community Psychology*, 1996, 24(3): 311-339.

重要产出的多元测量可以提供更宽泛的概念覆盖,并可以让一种测量弥补另一测量的缺陷。在统计上,可以将多元测量合并为单一的、更具活力的和正当的合成测量,这一测量优于任何一个单一测量。例如,在一个减少家庭人口生产的项目中,就可以合并测量家庭规模的改变、避孕措施的采用和对儿童的期望,以评估项目产出。即使测量的数量较小,详细阐述所有维度和变量对评估者从所有可选择办法中做出慎重的选择,也是有益的。

测量程序和特性

项目产出资料的基本来源相对较少——观察、记录、采访和问卷的回馈、标准化测试、物理测量工具和诸如此类的信息载体。当进行操作化时,这些信息就转变成测量标准,而操作化就是用一系列特殊的、系统的操作或程序进行产出测量。评估中的许多产出变量测量,会使用不同项目领域中已经建立和广为接受的程序与手段,对于较为长远的和与政策相关的项目产出更是如此。例如,在健康照顾项目中,以相对标准的方式进行发病率、死亡率和疾病或健康问题发生率的测量,测量方式视项目针对健康问题的本质不同而有所差异。例如,通常以标准化的成效测试和年级平均得分测量学术表现;而职业和劳工状况则一般通过官方统计局设计的测量方法进行评估。

对其他产出,可能也有多种现成的测量工具或程序,但是,哪个测量工具对

评估是最合适,通常少有共识。心理学的产出测量中,这个问题尤其突出,如对沮丧、自负、态度、认知能力和焦虑的测量,评估者的任务一般是从可行性的角度在有用的选项中做出合适的选择。因此在决策中,要考虑许多实践因素,如测量工具如何运用以及测量时间有多长。最重要的一个考虑是,现成的测量工具或程序与评估者希望测量的对象匹配如何,是否能够良好地反映出测量结果。专栏 7—B 阐明的说明,细致地描述将要测量的产出,有助于决策。专栏 7—D 的阐释也说明,如果评估者已经区分了与项目产出相关的不同维度,对测量工具或程序的选择也将很有帮助。

当使用现成的测量手段时,尤为重要的是确保所选择的指标能够代表重要的产出。合适的测量并非基于工具的名称或标签与重要产出相似。"同一"构造(例如,自负、环境态度)的不同测量手段通常具有相当不同的内容和理论导向(只有知道项目产出的细致描述,才能判断理论导向可能会或不会与重要项目产出特征相匹配)。

对于评估者来说,许多重要的项目产出既没有确定的测量标准也没有可供选择的现成测量工具。在这种情形下,评估者必须发展新的测量工具。不幸的是,这样做的话,时间和资源常常是一个重要的约束因素。一些特别的测量程序,例如从官方记录中吸收专门的相关信息,对于项目产出测量常常是可行的,十分简单且不需要进一步的实证。然而测量程序,例如问卷、态度量表、知识测试和系统观察的编码表,就不是这么简单了。建构这些测量工具,使它们以一致的方式测量评估者希望测量的内容,并不是一件容易的事情。因此,虽然有一些较成熟的测量程序,但就大量技术上的考虑,在有把握地使用这些测量标准之前,一般都要求进行大量的先期测试、分析、修订和确认(DeVellis, 2003;Nunnally and Berstein, 1994)。如果评估者在未完成以上步骤和核查的基础上就使用某个测量,最终的测量结果可能在表面上看起来是合理的,但对于精确评估项目产出来说,则并不必然是表现良好的。

当评估者没有机会以系统的和技术上恰当的方式来推进某一测量时,在做正式评估之前,核查基本测量手段的属性就显得尤为重要。的确,即使使用现成的测量工具和已被接受的工作程序来评估某些产出,首先确认不同的测量工具在特定项目情景中实施的优劣是明智之举,非常有必要区分不同测量程序工具及其相应特征。有三个测量工具特征是尤其需要给予关注的:信度、效度和敏感度。

信　度

测量的信度,是指当重复测量同一事物时,产生同样结果的可能性。那些重复测量的结果之间的变化组成了测量误差。例如,当同一封信在不同的场合下得出同样的"得分"(重量),这个邮局天平就是可信的。虽然没有完全可信的测量手段、分类方案或计算程序,但是不同类型的测量在变动范围内有可信度问题。使用标准设备对物理特征的测量例如身高和体重常常比心理上的特征测量

例如智力的 IQ 测试，更加具有可靠性。反过来，绩效测量例如标准的 IQ 测试，比依赖于回忆的测量例如消费品的家庭支出报告，具有更高的可信度。对于评估者来说，测量工具的性质是不可靠产生的主要根源，因为这些测量工具建立在参与者的书面或口头回应基础上。测试或测量情境之间的差异、观察者或访问者之间的差异和回答者的情绪波动都会带来不可靠性。

测量的不可靠性会冲淡和模糊真实的差异。如果一个真实有效的产出测量是不可信的，就会看起来没有实际的那么有效。评估者核查产出测量可信度的最直接方式是在同一情境下进行两次测量。从技术上讲，这种测试——再测试可信度，就是统计上所熟知的两组得分的积差相关（product moment correlation），其得分在 0.00 和 1.00 之间变动。然而，对于许多产出测量来说，这种检验很难操作，因为项目产出很可能在间隔较长的两次测量中发生变化。例如，如果希望了解学生喜欢学校的程度，那么即使是相同的问卷，在一个月后再问的时候可能会得到不同的答案。这不是因为测量工具不可信，而是因为干预事件使学生对学校的感觉发生了某些变化。另一方面，当产出测量涉及人们的回忆和印象时，与上次测量密切相关的再次测量也可能被扭曲，因为回答者会回忆起他们先前的答案而不是重新作答。在产出变化之前如果不可能重复测量，研究者往往通过同一时间内的多项测量来检验相似主题的一致性（即内在信度），来核对可信度。

对于评估者使用大量的现有测量工具而言，可以直接从其他研究或最初测量报告中获取可信度信息。然而，可信度会随着回答者和测量环境的变化而变化，所以，“在其他测量应用中可信的测量工具在评估活动中也是可信的”这一判断是不牢靠的。

没有关于可信度的普适的、不变而持久的规则。测量误差模糊有意义项目产出的程度，很大程度上取决于其在项目产出中所占份额的大小，我们将在第 10 章进一步讨论这个问题。就实践经验和偏好而言，研究者一般喜欢信度系数在 0.90 及其以上的测量，除了最小结果之外，这一范围可以使测量误差相对较小。然而，由于大量的产出测量是在项目评估的情境特征下展开的，所以，相对来说这是很高的标准。

效　度

测量效度问题比信度问题更加难以把握。测量**效度**（Validity）就是测量到希望测到内容的程度。例如，青少年被捕记录就提供了一个对行为不良的有效测量，这一测量仅在精确地反映多少青少年已经从事（可起诉的）犯罪时才是有效的。相反，警察的逮捕活动，则不是对被捕青少年不良行为的有效测量。

尽管效度概念及其重要性很容易理解，但检验某个测量对具体项目特征是否有效就比较困难了。假定运用产出测量进行评估，效度在很大程度上依赖于某个测量是否在有关项目方中获得共识，即认为测量是有效的。当对项目产出进行了充分而谨慎的描述时，确证测量结果的确代表了项目的预期产出可以提

供评估信度保证。使用多元测量也可以确保测量能接近实际产出。

从经验上看,效度常常来源于比较,即测量结果如果与人们的预期一致,就是有效的。例如,如果在运用某个测量工具时也运用替代测量工具,正如已经有评估者实践过的那样,那么测量的结果应该大体相同。同样,如果在不同的情境下运用相同的测量工具,那么测量的结果也应该不同。因此,一项环境态度测量应该能严格区分本地的塞拉(Sierra)俱乐部成员和远离道路污染的自行车组织成员。效度还可以通过与测量结果相关的效果或"预测"相关的其他产出而得到展示。例如,某个人的环境态度测量结果应该与回答者对政治候选人不同环境立场的赞许程度相联系。

敏感度

产出测量的首要目的是探明项目产出的变化或差异。为了很好地实现这一目的,测量工具应该对这些项目的效果敏感。测量的**敏感度**(Sensitivity)是指当测量的事物发生变化或有所差异时,能够测量到的产出变化程度。假设我们将体重作为减肥项目的产出。医师办公室使用的校准过的测量刻度虽然在几盎司范围内,也可以测量到体重减少的程度。相反,州际公路上测量卡车重量的刻度也是有效的和可信的,但对小于几百磅的差异却不敏感。如果测量工具对减肥项目中节食者的体重波动不敏感,那么测量结果就会很糟糕。

有两种主要方法可使项目评估中频繁用到的产出测量对项目产生的变化或差异变得不敏感。第一,测量针对的因素不是项目期望产生的变化因素。这样的测量会稀释甚至完全掩盖项目的作用。例如,小学生数学辅导项目在大部分学年里都关注分数和长除法问题。如果评估者选择教学大纲之外的数学问题进行测量,那么测量中就不仅仅包括了分数和长除法问题,还包括许多其他的数学问题。由此所得到的最终分数就会因为包括了大量的其他问题而掩盖分数和长除法教育的效果。因此,较为敏感的测量应该只有项目所教授的数学主题。

第二,如果产出测量仅仅用于诊断即探测个体差别时,那么对群体的差别变化可能会不敏感。这类测量的目的是希望测量个体具有某种特征的程度。大多数标准的心理学测量都是这种类型,例如人格测量、临床症状(压抑、焦虑等)测量、认知能力测量和态度量表。这些测量对探测在测量特征上的得分比较有效,这也是测量的目的所在,因此也是有益的,比方说评估需求或问题的严重性。然而,当用于一组项目参与者,而且在测量特征上参与者在参加项目前就存在巨大差异,参加项目之后虽然得分也可能发生巨大变化,但每个人进步的增量可能会消失在巨大的个体原有差异之中。从测量的立场来看,个体之间的差异对评估者希望测量的群组因项目活动所产生的变化构成了巨大的干扰,有时候甚至掩盖了项目活动所产生的效果。第 10 章将讨论评估者对这类测量不敏感的一些补偿方法。

在评估中,判断备选产出测量是否足够敏感的最好办法,是找到在已有的评估中充分敏感地测量了项目变化或差异的测量工具。当然,证明测量工具有效

的有力证据是在相似的评估中已经测量到了显著变化或差异。在证明测量工具的有效性时,还必须考虑先前评估研究的样本规模,因为样本规模影响到测量工具的测量能力。

研究测量工具敏感度的类似方法是将其运用到已知差异的群体中或已经变化的情境中,测量其反应程度。以前面提到的数学辅导项目为例,评估者可能希望看看学校系统每年进行标准数学测试的工具对项目产出是否充分敏感。假定辅导项目仅仅关注少量的数学主题,而标准测试的覆盖范围较大,那么工具的敏感性就值得怀疑。为了在使用前检验测量工具的敏感度,评估者可以首先对一个班级的学生在学习分数和长除法之前和之后进行测试。如果测量工具对变化足够敏感,那么这个工具就可以推广到数学辅导项目评估中。

产出测量的选择

正如讨论所示,在评估中为测量产出而选择最好的测量及方法工具就是评估研究的关键问题(Rossi, 1997)。我们力劝评估者投入必要的时间和资源,用于发展和检验合适的产出测量工具(专栏 7—F 给出了一个有教育意义的例子)。糟糕的产出测量工具不仅本身会导致差的效度,也不能恰当地实现评估项目的目标和目的。不可信的产出测量工具可能低估项目的有效性,并可能导致针对项目的不正确推论。简而言之,无效的或不可信的测量工具,由于产生误导性估计值,可能完全否定一项影响评估的价值。只有产出测量工具是有效的、可靠的和适度敏感的,影响的估计值才能是可信的。

专栏 7—F 对无家可归精神病患者采用自报测量方法的信度和效度

对无家可归精神病患者项目的评估,主要依赖于自报测量方法。但是这种测量方法的信度和效度如何?尤其对有精神问题的人?一组评估者将一种测量工具引入了对无家可归精神病患者案例管理服务的评估中。他们把焦点放在了对精神病症状、物质滥用和服务利用资料的自报测量方法上。

精神病症状:进行简单症状目录(BSI)自报是精神病症状评估中的主要测量方法。通过5批数据的搜集,对内在信度进行了测量,在焦虑、沮丧、仇恨和躯体化尺度方面表现出很高的信度(0.76~0.86);但是,对精神病症状而言,表现的信度较低(0.65~0.67)。为了获得这些测量的效度证据,计算了其与简明精神病分级表(BPRS)(由硕士水平的精神病学家和社会工作者对对象进行的评定分级)中可比尺度的相关关系。通过5批数据采集,就焦虑、沮丧、仇恨和躯体化而言,其相关性表现为中度一致(0.40~0.60),但就精神病症状而言,相关性较差(-0.01~0.22)。

物质滥用:该测量方法是用嗜好严重指数(ASI)估计顾客对酒精和其他物质滥用的程度。为了有效起见,访问者根据同样的嗜好严重尺度指数将病人按对酒精和其他物质滥用的需求分级。通过5级测量,与酒精嗜好的相关性为0.44~0.66,表现为中度一致,与毒品的相关性为0.47~0.63。自报比访问对服务需求的报告数要少。

项目接触和服务利用:病人报告他们与项目接触的情况以及他们是否接受了14项特别服务中的任何服务。通过与两批测量的案例管理报告进行比较,对这些报告的效度进行了检验。实际上,一致性依内容的变化而变化。就与项目的接触、支撑性服务和特别资源领域(立法、住房、金融、就

业、保健、医疗)来说,相关系数最高(0.40~0.70)。对精神健康、物质滥用和生活技能培训服务而言,一致性相当低。大多数的不一致是案例管理员报告了某种服务,而病人没有报告。

评估者们由此得出结论:对无家可归精神病患者用自报测量方法是正当的,但警告:评估者不应仅仅依赖自报测量方法评估精神病症状或精神健康、物质滥用的服务利用,因为病人在这些领域给出了显著过低的估计值。

资料来源:Robert J. Calsyn, Cary A. Morse, W. Dean Klinkenberg, and Michael L. Trusty, "Reliability and Validity of Self-Report Data of Homeless Mentally Ill Individuals," *Evaluation and Program Planning*, 1997, 20(1): 47-54.

督导项目产出

随着对重要项目产出充分测量程序的形成,评估者或项目管理者发展了多种方法来了解项目产产出。最简单的方法是产出督导,即我们已经在第6章界定过的对项目意欲改善的社会状况的"持续测量"和指标报告。正如第6章描述的那样,这与项目督导相似,不同的是,有规律地搜集和评估的信息涉及项目产出而不仅仅是项目过程和执行。例如,一个工作培训项目的产出督导可能包括项目完成6个月后打电话询问参加者是否就业?如果就业,进一步询问从事什么工作以及工资是多少(对产出督导的讨论,详见 Affholter, 1994; Hatry, 1999)。

产出督导要求有识别重要项目产出的指标,指标的测量具有操作性,能反映项目的绩效状况。后面的那个要求尤其难以做到。正如本章前面所讨论到的,对产出进行简单测量所提供的信息仅仅是产出的状况或水平,例如贫困儿童的数量、药物滥用的流行程度、失业比率或小学生的阅读技能。困难所在是要识别项目对象状况的变化,尤其要识别与项目效果或影响的方向相关的变化。

正如前面提到的那样,测量困难的可能来源是——社会状况通常受到项目之外多种因素的影响。因此,贫困率、药物滥用、阅读得分诸如此类产出可能因为经济、社会趋势和其他项目、政治影响等相关大量因素的作用而变化。在这些条件下,我们应该明白,构建可以把其他因素的影响与项目效果相剥离的产出测量指标并不是一件容易的事。以令人信服的方式将项目效果从其他具有相似效果的影响中剥离出来,需要专门的影响评估技巧,第8和第9章将对此进行讨论。

综上所述,产出督导通常在合理的时间进度内提供项目效果的有用的和相对低成本的信息。一项影响评估可能需要几年时间才能完成,而产出督导的相应结果则可能在数月内就可以利用。此外,影响评估的成本通常比产出督导大得多。但是,由于产出督导也有局限性,因此主要被当作项目反馈以协助项目经理更好地管理和提高项目干预技巧之用,而不是用来评估有益于社会状况变化的项目效果。这里,以酒精中毒治疗项目的产出督导作为例证。80%的项目对象在项目结束后几个月就不再饮酒,比20%放弃饮酒的产出具有更强的一致性。

当然，两个产出都不足以构成真实的项目效果，因为饮酒的程度和其他可能超越项目自身影响的因素，都将影响戒酒的测试水平。一个好的督导方案应该包括问题最初的严重性、对其他重要影响的了解程度和其他相关因素的指标。由于正式影响评估的缺乏，合理地解释和进行指标之间的比较，尤其是项目试图提高绩效的指标的趋势比较，都可以提供涉及项目有效性的有价值的指示。

产出督导的指标

产出督导的指标应该尽可能回应项目效果。例如，产出指标应该仅仅测量实际接受项目服务的目标人群。这就意味着项目地理范围内容易利用的社会指标如人口普查地域、邮递区号、自治市，如果包括相当数量没有实际接受项目服务的对象，就不是好的产出督导指标。也意味着，对于尚未完成整个项目过程的项目参与者而言，一系列服务后果尚未显现，故应排除在指标计算之外。这不是说辍学学生比例不是项目执行测量的重要指标，而是说应该将其作为服务利用问题而不是产出问题来评估。

除了影响评估，最具解说力的产出指标是那些仅包括了项目影响的变量。如果这些变量是项目预期的主要产出，则也有利于产出督导。例如，城市街道清洁项目的目标是清洁城市街道上的垃圾、树叶之类的杂物。独立观察者认为，街道清洁照片可以为评估项目有效性提供情报。只要没有飓风将所有的垃圾吹进临近的国家，清洁街道的项目就是有意义的。

最易于直接与项目活动连接起来的产出指标是对象满意度，即所谓的人类服务项目中的消费者满意度。项目受益者的直接排序，提供了一种评估产出的形式。另外，创造项目参与者之间互动的、满意的感觉，也是产出的一种形式，尽管它本身不会必然提高参与者的生活水平。更多的信息来自参与者的报告——特殊收益是否来自于项目传递的服务（见专栏7—G）。不过，这些指标的局限性是项目参与者可能不能经常识别或了解项目收益，正如在吸毒成瘾者的例子中鼓励成瘾者使用洁净针具一样。相应地，参与者可能有收益，但并不情愿对收益估计过高，正如征求老年人是否需要护士上门服务的案例所显示的那样。

专栏7—G 与具体收益相关的对象满意度调查问题

对象满意度调查主要关注项目服务满意程度。项目对象感到满意是项目产出的一种，单有这一点很难说明对象对特定项目收益感到满意。为了调查超出服务问题的对象满意状态，必须询问服务结果的满意度，即对服务可能产生的特定变化的满意度。马丁和凯特勒（Martin, Kettner）建议增加如下一些问题以执行对象满意度调查：

服务：信息和建议

问题：信息和建议项目对于你评估服务需求有帮助吗？

服务：家庭—送餐

问题：家庭—送餐项目对于你维持健康和营养有益吗？

服务：咨询

问题:咨询项目对于你处理日常生活压力有帮助吗?

资料来源:Lawrence L. Martin and Peter M. Kettner, *Measuring the Performance of Human Service Programs* (Thousand Oaks, CA: Sage, 1996), p.97.

产出督导的缺陷

因为项目试图影响社会状况的动态性、产出指标的局限性和项目机构的压力,项目产出督导具有许多缺陷。因此,一旦项目产出指标可以被用作项目决策者有价值的信息来源,在使用上就必须小心谨慎。

一个重要的考虑是,项目投资者和有影响的决策者密切关注的任何产出指标,将不可避免地受到项目职员和管理者的重视。如果产出指标是不合适的或者未能覆盖所有重要项目产出,那么指标反映的绩效提升将扭曲项目活动的真实效果。例如,阿弗尔特(Affholter, 1994)描述了一个多重问题的情形:某国家将许可的养育者家庭作为新增儿童空间的指标。即使他们缺乏与孩子们共处所必需的技巧,工人们还是精神旺盛地增加新成员和建立新家。这样一来,指标持续上升,但合适养育者家庭中儿童的实际增长空间却并没有提高。在教育上,这叫做"为测试而教学"。相反,好的产出指标必须是"为教学而测试"。

一个相关的问题是"指标的腐败性",即一种很平常的倾向:在评估中,测量者为了让指标使其绩效比实际上看起来要好,只要有机会,就可能蒙混和夸大指标测量值。例如,如果在项目实施后把参与者获得工作的比率作为主要的产出指标,由于项目职员所面对的压力,在后面的电话随访了解项目参与者获得工作的情况时,即使他们很正直且做出了合理的努力,其结果很可能是尚不明确的数量多于失业的数量。如果是在项目内收集这类信息,使用严谨的程序和令人信服的方式核实测量结果就尤为重要。

另一个潜在的问题与产出指标的解释有关。假定存在一系列可能影响指标的因素(除项目实施影响之外),即使具有适当的相关情景,其解释也可能令人误解和难于理解。产出指标通常必须与其他建立在产出指标基础上的、可以提供产出比较或解释的信息一起,才能够为解释提供合适的情境。我们认为这类信息对于接下来的产出数据解释十分有用。

解释产出数据

如果没有项目对象变化的信息、相关人口和经济趋势等信息相配合,那么,对作为常规产出督导的一部分而收集的产出数据进行解释将十分困难。例如,如果我们知道项目参与者失业问题的严重性和当地经济中职业空缺的范围,作为项目绩效的一个指标,工作安置比例就可以得到更为准确的解释。当项目对象具有较少技能和长期失业,并面临着具有较少职业空缺的经济发展环境,那么,低的安置比例也许不能对项目绩效做出有效的反映。

同样,如果有项目过程和服务利用信息的支持,产出数据往往具有更强的说服力。即使完成培训的对象有较高的工作安置比率,但是如果完成培训的比率并不高,也不能说明项目有好的效果。良好工作安置比例的实现可能是因为所有具有严重问题的对象都已退出,只剩下需要项目安置的且易于安置的对象。合并过程和利用信息在解释产出指标中十分重要,尤其是当比较不同的单位、地点或项目时,价值更大。仅仅因为一个项目比其他项目在产出指标上低,而不考虑是否面对更多的棘手对象、较低的放弃率,或其他情有可原的因素,就做出项目效果的否定判断是不准确和不公平的。

对产出督导数据解释同等重要的是提供判断标准的逻辑框架。这些标准用来依据既有数据判断哪些因素使项目更好或更坏。在督导数据可用的条件下,一个有用的框架是通过项目实施前和实施后的数据变化来进行产出测量。例如,仅仅知道在经过6个月的职业培训之后有40%的人找到了工作比之知道在培训之前项目参与者中有90%的人没有工作所能提供的信息要少。获得产出指标的一个途径是界定项目参与者"成功的门槛",并说明多少人在接受服务之后从那个"门槛"之下移到了之上。因此,如果门槛被界定为"持续全职工作6个月",需要说明的就是在项目实施前全职工作少于6个月和项目实施后全职工作多于6个月的比例。

虽然这类简单的前后(之前和之后)比较不需要成为常规产出督导的一部分,但却可以作为产出评估的一部分。正如我们已经注意到的,这一设计的主要的缺点是:测量到的前后差异不能完全归结于项目效果和影响,因为干预期间的其他运作过程可能影响前后之间的差异。例如,人们选择参加工作培训的主要原因之一是他们正处于失业状态并且在获取雇用机会上有困难。因此,他们进入项目时是处于一个低点,一些人可能并没有将工作与参与项目联系起来。因而,雇用前后比较与项目效果之间并没有必然的联系。

两段时间之间的其他趋势也会影响前后变化。一个降低犯罪的项目,如果与增加管束和治安的政治努力不谋而合,就可能更为有效。干扰因素则会扭曲前后比较:如果一个职业培训项目处于较长的失业率上升和经济低迷时期,将可能呈现出无效。因而,一般而言,前后比较可能为项目管理者的产出督导提供部分有用的反馈,但通常不会提供可信的项目影响信息。事实上,极少有没有其他干扰或趋势影响前后比较差异的情况,专栏7—H提供了这种情形的一个例子。

专栏7—H 令人信服的一项减少低收入者住房中铅水平的前—后产出设计

铅对儿童尤为有害,会阻止儿童的行为发育,降低其智力、引起听力丧失,妨碍其重要的生物、生命功能。贫困儿童遭受铅毒害的风险更大,因为适于低收入租客的一般都是老房子,更可能使用含铅颜料粉刷,并离其他铅污染源较近。室内含铅油漆产生的铅微粒可以通过手—嘴活动被儿童摄入。更甚者,风吹或车行扬起的灰尘都可能被铅油漆污染,或者路边的泥土含有1980年以前使用含铅石油造成的铅污染。

为降低低收入郊区住宅的铅灰尘水平,联合公共—私人的努力在巴尔的摩(Baltimore)发起建

立了社区铅教育和减少公司(CLEAR 公司)。CLEAR 公司成员负责清理、修缮,使住房达到铅安全水平,教育居民预防铅中毒的技巧,鼓励居民通过特殊的清理保持灰尘的低含铅水平。为判断 CLEAR 公司在降低郊区住宅灰尘含铅水平成功的程度,在铅控努力 6 个月后,CLEAR 公司成员收集之前和之后灰尘含铅样本。在43 所住宅中的每一所,都从 4 个位置采集样本——地板、窗台、窗井和地毯——并送到实验室进行分析。

地板、窗台和地毯前后灰尘铅水平测量之间具有显著的差异。6 个月之后,地板和窗井的灰尘铅水平进一步显著下降,窗台灰尘铅水平的下降则相对不太显著。

由于没有使用控制组,除了 CLEAR 公司项目之外的其他因素也可能影响灰尘铅水平的降低。除了直接相关的影响,谨慎地看,还可能有季节性的影响,也可能有针对同一住宅的其他干预项目的影响,由于没有可用的证据,因此,基本上可以认为差异较显著地来源于项目干预。因此,评估者得出结论,CLEAR 公司项目在降低住宅铅水平上是有效的。

资料来源:Jonathan P. Duckart, " An Evaluation of the Baltimore Community Lead Education and Reduction Corps(CLEARCorps) Program," *Evaluation Review*,1998, 22(3):373-402.

项目产出变量测量或其他变量测量值的变化信息,必须由项目管理者、项目方或专家来判断并进行解释,因为与他们对项目的或好或坏的预期和判定相关。这些判断在两个极端点上最为简单——当产出由于项目之外的原因而比可能发生的要积极时,或者非常消极以至于对失败很难解释。

假设在两个月的拖斗式卡车司机职业培训项目实施之后,90% 以上的参与者(从那些不具备这些技巧的人中挑选出来)有资格获得合适的驾驶执照。评估发现表明:项目在传授职业技巧上是成功的——尽管初看起来这么大比例先前没有技术而希望成为拖车式卡车司机的人在两个月时间内获得驾驶执照是不可能。同样,我们也可以有另一个极端的判断,即如果所有参与者都没有通过驾照考试,这个项目就是无效的。

当然,在现实中,观测到的结果很可能更加模糊——仅 30% 的人一次通过。根据这一更为典型的结果,很难判断和推测没有接受培训的比较组是否会做得同样好。专家判断可能注意到这样一些条件——例如,可能会要求熟悉成人职业教育和熟悉该领域干预项目典型产出的人运用专业知识和训练,去判断 30% 的测量结果是否能代表项目在目标人群中的成功。显然,这些判断的有用性和有效性以及使用性价值,主要依赖于评判者在这一项目领域内的专有技术和知识水平。

很可能,这类产出测量值可以与类似项目的产出进行比较,这一过程通常被叫做"基准化(benchmarking)"(Keehley et al. ,1996)。尤其是,当具有某一特定产出的项目与另一个极其有效的项目进行比较时,这一原理显露无遗。当然,只有在其他条件可比的前提下,这些比较的结果对于评估才是有意义的。不过,在大多数情况下,都很难满足这一标准。

小　结

- 设计项目是为了以积极的方式影响某些问题或需求状态。评估者通过测量产出、目标人群状况或项目预期改变的社会状况,来评定项目对某一方面状况的改善程度。
- 既然项目产出会受到项目外独立事件和经验事实的影响,那么,产出水平的变化就不能直接解释为项目效果。
- 界定项目产出,需要项目方的信息反馈,需要进行项目文献回顾,以及在项目逻辑基础上对影响理论的明晰化。评估者应该考虑相关的已有研究和项目在实施过程中的意外产出。
- 为了获得可信的项目产出,产出测量工具应该是可靠的、有效的和充分敏感的,通过对项目预期产出大小的排序,核查项目产出水平的变化。另外,可以使用多元测量标准或多个产出测量变量来反映多元结果、校正单个或更多个测量中可能存在的缺陷,这种做法在很多时候都是明智的。
- 产出督导通过提供及时而相对低成本的研究发现,可以指引项目进一步调整和改善,从而为项目管理者和其他项目方服务。有效的产出督导,要求对测量指标进行仔细选择、对产出数据进行谨慎解释。
- 使用上述方法来解释产出测量值和项目产出的变化是不太容易的。可靠的解释需要考虑项目环境、项目实施期间发生的事件和目标群体所经历的历时性自然变化。另外,解释通常有赖于专家对项目绩效好坏的判断;同时,与其他项目(标准)的比较也是十分有益的。

基本概念

影响(Impact):参见“项目效果”。

产出(Outcome):目标群体的状态或者项目预期达到的社会状况改善。

产出变化(Outcome change):不同时间点项目产出水平之间的差异。参见“产出水平”。

产出水平(Outcome level):某一时间点上的项目产出状况。参见“结果”。

项目效果(Program effect):在项目的产出变化中,在其他影响得到控制或排除的条件下,可以完全归因于项目干预作用的变化,也称作项目影响。参见“产出变化”。

信度(Reliability):使用某种测量工具重复测量同一事物产生相同结果的程度。

敏感度(Sensitivity):当被测事物发生变化或存在差异时,测量值发生变化的程度。

效度(Validity):测量工具实际可以测到特定变化的程度。

8 项目影响评估——随机实地实验

影响评估就是确定干预是否在实践意义上产生了预期效果。这类评估不可能以十足的确信度去进行,而只能得到不同程度的似真性。其一般的原则是:研究设计越严格,干预效果估计值越真实。

影响评估设计需要应对两方面具有挑战性的压力:其一,应充分严格地从事评估,以便能获得相对可靠的结论;其二,对时间、金钱、合作和参与者保护的实际考虑,限制了所能采纳的设计选择和方法及程序。

通过将项目参与者的产出信息与假定他们没有参加项目的产出信息进行比较,评估者对社会项目效果进行估计。本章将讨论达成这一目标最完善的研究设计——随机实地实验。随机实验基于随机分配接受干预的实验组和未接受干预的对照组之间的比较。尽管实践考虑会限制随机实地实验在某些项目情境下的使用,但是,评估者仍需要熟悉它们。随机实验的逻辑是所有影响评估设计和数据分析的基础。

影响评估用来确定项目对预期产出产生了什么样的效果以及是否存在重要的意外效果。正如第7章谈到的，项目效果或影响，是指由项目引起的目标群体或社会条件的变化，即这种变化不是发生在项目影响之外。因此，弄清楚项目效果与分析项目之外某些特殊因素的效果是同一个问题。

在社会科学中，因果关系基本上建立在概率基础之上。因此，如果说“A导致B”就意味着如果引进A，B产生的概率大于我们不引进A时的相应概率。这一表述并不暗示B通常由A引发，也不是说仅当A首先发生时B才会发生。举例说明，一项工作培训计划的目的是减少失业：如果成功，它将增加参与者最终被雇用的概率，但并不一定给每个参与者都带来工作。找到工作的可能性还与培训项目影响之外的因素相关，例如，社区中的经济条件。相对地，一些项目参与者可能没有项目的协助也能找到工作。

因此，影响评估中的主要问题是，在没有干预或在某些案例采用替代干预的情况下，产生的预期效果是否会超过进行干预时的对应效果。本章，我们将讨论解决这类问题最常用的研究设计——随机实地实验。我们从对影响评估的一些基本考虑开始展开讨论。

开展影响评估的时间选择

影响评估与社会项目发展历程中的许多节点都可能相关。在对政策进行阐明的阶段，示范性试验项目可能需要进行影响评估，以判断提出的项目是否将真正地产生预期效果。而当一个项目得到批准后，这种影响评估通常会在有限的地点同时展开。在这种情况下，影响评估的结果可能是正面的：即项目在扩展其覆盖面之前，已经取得了一定预期效果。在许多情况下，革新性项目的主办者，例如私人基金，会在有限的范围内贯彻项目干预并进行影响评估，如果干预效果得到证明，就可以提高政府机构对项目的采纳。

实施中的项目也经常需要进行影响评估。在一些情况下，项目需要修改和完善以提高它们的效率或适应修正后的项目目标。如果变化是显著的，那么，修正过的项目就可以保证实施影响评估的合理性，因为它实质上成为了一个新的项目。但是，许多稳定的、已固定化的周期性影响评估项目也是合理的。例如，针对某些高成本的治疗方式，基本做法就是不断评价它们的功效，并将它们与其他处理相同问题的方式进行比较。在很多情况下，长期项目都需要定期进行评估，因为“日落立法”①要求为政府资助的项目提供绩效证明，以获取下一步的项目资金；或者，作为项目回应抨击的一种方式，影响评估使项目免于替代性干预方法或其他公共基金支持者的抨击。

---

① 各项计划都规定期限，并须经过特定的评估，否则不予拨付运作基金。在法律上，所谓“日落立法”是指授予行政机关一定立法权，经过一段时间后，非经再授权，则行政机关的行政立法权自行失效。——译者注

不论在何种情况下进行影响评估,都需要一些使其有意义的先决条件。首先,影响评估建立在早期其他形式评估基础之上。在进行项目的影响评估之前,评估者需要评估项目理论和项目过程。项目理论的评估应该指出项目的目标表述十分明确,以使具体的预期效果成为可能——这是针对项目效果进行影响评估的必要先决条件。而且,合理的假设是,项目活动产生了这些效果。项目过程评估则应该显示出,项目干预得到了充分实施,从而有机会对干预对象产生预期效果。评估不合理的、不可测量的项目产出或者是尚未充分实施的项目,是一种浪费时间、精力和资源的做法,很不值得。以上诸多考虑的一个重要意义是:只有当项目干预已经介入了足够长的时间,已经消除了基本实施问题时,才应该来考虑项目干预的影响。

更为严格的影响评估形式会涉及重要技术和管理的挑战,认识到这一点非常重要。社会项目的目标群体通常是很难接触的个人或家庭,或者说,研究者很难轻松地从他们那里获得项目产出资料和跟踪数据。另外,比较可信的影响评估设计要求有技术和时间这两个因素的保证。最后,正如我们将在第 12 章详细论述的那样,评估研究也需要政治背景的支持。评估者必须不断地培养项目职员和目标参与者之间的协作,以引导影响评估,同时,还须与要求产生及时、明确的发现这种内在压力相"斗争"。因此,在进行影响评估之前,评估者应该适当考虑对项目环境中可利用资源和信息的判断是否充分。项目各方会经常询问影响评估的结果,因为他们对项目是否产生预期收益感兴趣,但是,他们可能并不会认为项目的某些先决条件和研究资源对于以可靠的方式完成这一任务是必要的。

影响评估的关键概念

本质而言,所有影响评估都是比较性的。要确定影响的大小,就需要严格且可行地将接受某种干预的对象与接受其他干预的同等对象进行结果比较。在实践中,通常将项目参与者与具有其他经历的类似个体进行比较。接受"其他干预"的对象可能有一个或一个以上的群体,"其他干预"意味着接受某些替代性服务或者仅仅是不受特定干预影响而已。用于比较的"同等对象"可以通过各种不同的方式选择,或者选取不同类属的几个群体进行比较,或者根据所要核查的项目产出观测值与没有干预时同样对象的先测值进行比较。

理想而言,用于比较的对象应该在项目干预之外的所有方面都是相同的。近似于这一理念的几个替代性的(但不是相互排斥的)方法在有效性上有所差异。不过,它们都涉及控制组的建立——确定控制条件以保证环境中的某些对象没有受到评估干预的影响。可用的选择对象的方法是不均衡的:本质上,在项目影响估计方面,一些方法比另一些方法更可靠。各种方法的成本和所需技术、技巧水平也不相同。其他条件相同的情况下,产生更有效评估结果的方法通常要求有更娴熟的评估技巧、更多的时间投入和更高的成本。一般而言,可以把影

响评估的设计方法分为两类，这也是我们接下来将要讨论的主要内容。

实验型研究设计与准实验型研究设计

我们对影响评估可选择方法讨论的基础是：最有效地确定干预效果的方式是**随机实地实验**（Randomized field experiment），通常又被叫作评估项目效果的"黄金标准"研究设计。随机实验方法的实验室描述无疑是最常见和最基本的。参与者被随机分成至少两组。一组为**控制组**（Control group），不受干预或干预无效；另一组则为**干预组**（Intervention group），接受项目干预。同时观察和比较两组的产出，所得差异就可以归因于项目干预效果。

随机实地实验的控制条件建立在相似构成基础上。研究对象被随机分派进干预组和控制组，干预组接受项目干预，控制组则不接受干预。可能同时有几个干预组，每组接受不同的干预或同一干预的不同量；也可能同时有几个控制组，对每组进行不同的控制，例如，无干预、"安慰剂"干预以及项目干预情境中其他正常可用的控制类型。

所有剩下的其他影响评估设计共同组成了非随机的**准实验**（Quasi-experiment）方法，其目标是将参与项目者（干预组）与非参与者（控制组）作比较，这些非参与者在主要变量上与参与者相似。因为缺乏随机分组所需的基本条件，所以这些方法被叫做准实验设计。下一章内容将讨论在影响评估中建立非随机控制组的主要方式。

使用非随机控制设计一般比执行良好的随机实地实验，在产出方面具有更低可信度。因此，从项目产出评估有效性的立场上看，随机实地实验通常是影响评估方式的最佳选择。然而，当实施随机实验不实际或不可能时，准实验方法对于影响评估也是有用的。

评估项目效果的不同研究设计的优势和弱点，这些设计方案下实施影响评估并分析结果数据的技术细节，都是评估的重要话题（这方面的经典文章请参见：Campbell & Stanley 1966, Cook & Campbell 1979；评估者可以在以下的一些论述中发现更多有用的信息：Shadish，Cook, & Campbell, 2002 和 Mohr, 1995）。

"完美"与"足够好"的影响评估

由于某些原因，评估者会经常面对这一问题：实现"真正最好的"影响评估设计非常困难。首先，在技术上最好的设计，因为在干预群体或目标对象范围内不适用而只能被放弃。例如，在伦理和实践上与人性问题相关的随机实验就可能被评估者弃用，而换用不太严格的设计。第二，时间和资源的有限性总是会限制设计方法的选择。第三，对最好评估设计的证明（经常是花费最大的过程）会随着项目干预重要性和干预产出的预期利用而不同。同样，一个重要的项目（因为试图改善一个非常严重的条件或采纳一种有争议的干预而具有重要性的项目）应该进行比其他项目更为严格的评估。而在其他极端情况下，某些实验项目可能根本就不用进行影响评估。

为了确定具体评估的最合适设计,我们的立场是:评估者必须了解设计选择的范围。选择总要涉及权衡,没有哪个设计总是最好的,并能作为"黄金标准"而被广泛应用。相反,在构筑研究设计时,我们提倡用"足够好"的原则。简单来说,"足够好"原则就是评估者应该在考虑结果潜在重要性、每一设计的实用性和可行性以及所选设计产生有用及可靠结果的可能性以后,从方法论角度选择最可能的设计方案。在本章剩余部分,我们将首先关注在方法论上最严谨的影响评估设计——随机实地实验。

随机实地实验

如前所述,项目产出或影响可被概念化为接受特殊干预的目标群体和未受干预的"对等群体"在产出方面的差异。如果两组完全对等,并将受到相同程度的项目外因素变化产生的影响;那么,产出之间的任何差异就应该代表着项目影响。影响评估的目的,尤其是随机实地实验的目的,是分离和测量任何此类差异。

以这种方式评估项目影响的最重要因素,就是形成一个控制组(其成员不参与项目,除此之外,其他方面与参与者相同)。由于这些目的,"对等"意味着以下这些意义:

- 构成一致:按照与项目和结果有关的特性,干预组和控制组由同样的人员或单位组成。
- 趋向一致:在没有干预的条件下,干预组和控制组面对同样的项目环境,且具有同等可能性达到既定的产出状况。
- 经历一致:在整个观测过程中,干预组和控制组经历同样的与时间有关的过程:成熟、长期趋向、干扰事件等。

尽管在理论上,可以通过将研究对象两两对等地分别与干预组和控制组匹配达到完全相同,但这在项目评估中几乎是不可能的。没有两个个体、家庭或其他单位在所有方面都是相同的。所幸的是,一对一的相同并不是必要的。干预组和控制组只需在群体意义上与评估项目的预期产出有关的方面一样即可。对干预组和控制组成员在出生地、年龄方面不同的项目进行影响评估是根本没有关系的,只要这种差异不影响到产出变量。另一方面,任何与研究结果相关的干预组和控制组之间的差异,都将引起项目效果估计值上的偏差。

使用随机选择建立对等性

获得干预组和控制组之间对等性的最好方式,就是使用**随机选择**将对象群体的成员分配给这两组。随机化是一种确定个体(或单位)是接受项目干预或替代性控制条件的程序。需要指出的是,这种意义上的"随机"并不意味着偶然或反复无常。相反,将对象随机分配给干预组和控制组需要非常谨慎,以保证对象群体中的每个单位都有同样的机会被选择进任意组。

要创造一个真正的随机配置，评估者就必须使用以概率为基础的精确分派程序，例如，随机数字表、轮盘赌、掷骰子等。为方便起见，研究者经常使用随机数字表。随机数字表在大多数初级统计或样本抽样的教科书中都有附带，许多计算机统计软件包也包含有产生随机数字的子程序。影响评估中，对每一个参与者进行分组的基本步骤是确保选择完全基于下一个随机结果，例如随机数表中的下一个数字（例如，奇数或偶数）（对如何实施随机化的讨论，请参见 Boruch，1997；Boruch & Wothke，1985）。

因为干预组和控制组之间的差别是偶然的，所以，除非这种机会波动不定，否则，与产生项目产出的干预相对照的任何其他影响在两组对象上都是相等的。就像将一把钱币抛向空中，正面和背面朝上的几率相等一样。例如，如果让人们自由选择，由于随机化，那么选择干预组和控制组的概率应该是相等的。所以，两个组将有同样比例的人趋向于接受干预。

当然，即使采用随机分组，干预组和控制组绝不会完全可比。例如，可能碰巧进入对照组的妇女比实验组多。但是，如果反复做随机安排，这些变化就会平均化到零。用合适的统计模型，就可以计算出在一系列随机化过程中，对具有任何特性、规模差异的对象进行随机化的次数。因此，干预组和控制组在产出方面的任何差异，都可与基于机会（即随机过程）的期望值进行比较。这样，统计显著性检验容许判断某种差异是否仅仅因偶然性而发生；或者不是由偶然性引起，从而，更能代表干预的效果。因为，在运作良好的实验中，项目干预是干预组和控制组之间的、除偶然性外的唯一差异，所以，这种评价成了辨别项目效果存在的基础。进行这种计算的统计程序是相当简单的，可以在任何关于统计推论的教科书中找到。

偶然性和统计显著性检验的一个重要暗示是，影响评估需要有较多的案例作为基础。随机分配到干预组和控制组的案例数量越大，那么，组与组之间就越具有统计上的对等性。其原因类似于将 1 000 枚硬币抛向空中所得到的偏离正反面向上次数均等的可能性要比仅抛两枚硬币时的相应概率低。如果每一组评估对象仅有一个或少量的研究单位，就不大可能妥善地进行影响评估，因为其中任何数量个体间的不均衡将会导致统计上的显著差异。这一问题以及与之相关的其他内容将在第 10 章进行更加全面的讨论。

分析单位

影响评估中所采用的产出测量单位被称作**分析单位**。影响评估的分析单位并不必然是个人。社会项目也许要涉及各种不同的对象，包括个人、家庭、邻里、社区、学校和商业公司等组织，以及从县到整个国家的政治权力机构。尽管实地实验的成本和难度可能会随单位的大小和复杂性的增大而增加，但是，从一种单位到另一种单位，影响评估的逻辑思想是不变的。例如，执行一项实地实验并对 200 名学生搜集数据，通常比对同样数量的班级或学校进行比较评估要容易得多，而且评估成本低得多。

分析单位的选择由干预的性质和被服务送达的对象单位决定。旨在通过大

宗拨款给当地市级机构来影响社区的项目,其研究单位就是市级机构。请注意,在这种情况下,每一个市级机构将构成分析的一个单位。因此,通过对比两个市级机构进行的拨款的影响评估,其样本大小是2(这对许多统计目的而言是完全不够的),尽管可以在每一个社区内分别进行大量的个体观测。

试图设计影响评估的评估者应该把被指定的单位作为研究干预的对象看待,并指定为分析对象。在大多数情况下,确定分析对象是很明确的;在其他一些情况下,评估者则需要仔细揣摩项目设计者的意图。在另一些情况下,干预要涉及一种以上的对象:如,住房补贴项目可能旨在提高各贫困家庭的居住水平和当地住房储备。这里,评估者可能希望设计一种由当地社区内家庭样本组成的影响评估。这样一个设计将涵盖两类分析单位,以评估项目活动对单个家庭住房以及社区住房储备的影响。这种多层级设计与仅有一个分析单位类型但却包含更为复杂统计分析的实地实验遵循同样的逻辑。

随机实验的逻辑

专栏8—A列出了一个简单的"干预前—后"随机控制实验的示意图,阐明了随机实验进行项目效果评估的逻辑。在项目实施前后,分别计算干预组和控制组产出变量上的平均值变化。如果组与组之间(除项目参与者之外)对等性的假设是成立的,控制组变化的数量就代表着在没有接受项目干预的成员身上所可能会发生的产出数量。当把这个数值从干预组产出变量的变化值中减去时,剩下的变化量就直接意味着项目效果。

但是,干预组和控制组之间的差异(专栏8—A中的I减C),也可以来自最初随机样本配置时的统计因素。因而,干预组和控制组之间平均结果的具体数字差异并不能简单地被解释作项目效果。我们必须提供一种合适的统计显著性检验,以判断某种程度上的差异是否由偶然性所致。适用于这种情形的一些常规统计检验包括t检验、方差分析和协方差分析(用先测值作为协变量)。

专栏8—A的示意表达式将效果表示为干预前和干预后产出的差异。不过,对某些类型的产出,我们不可能确定干预前的测量值。例如,在一个预防性的项目中,在项目服务送达之前,通常不会出现体现预防对象的相应产出。我们可以看一下预防青少年怀孕的一个项目,虽然也针对那些尚未怀孕的青少年,但是"怀孕"却是项目意欲影响的主要产出。同样,帮助高中生进入大学的项目,其主要产出也只能在干预之后才看到。但是,如果条件允许的话,同时获得干预前和干预后的测量值,往往具有统计上的优势。例如,当进行干预时,获得每一个初始点的测量值,将会使项目效果评估的估计值更加精确。

专栏8—A 随机实验示意表达式

	产出测量		
	项目前	项目后	差值
干预组	I1	I2	I = I2 - I1
控制组	C1	C2	C = C2 - C1

项目效果 = I - C,其中:

I1,C1 = 项目开始前,干预组和控制组各自的产出变量测量值。

I2,C2 = 项目完成后,干预组和控制组各自的产出变量测量值。

I, C = 干预组和控制组各自的总产出。

影响评估中随机实验的几个例子

有几个例子可以说明应用于影响评估中的随机实地实验方法,也可以解决现实生活中遇到的一些困难。专栏 8—B 描述了一项旨在评价改变学生不良饮食习惯干预效果的随机实验。这里,有几个实验特性是有关的,需要注意:第一,分析单位是学校,而不是学生个体;然后,将所有学校分配到干预组或控制组。第二,多元产出测量包括了多个营养干预目标,另一个重要的方面是用统计检验帮助判断(干预组的能量摄取总量和从脂肪中摄取能量较少时)净效果是否只是随机误差所致。

专栏 8—C 描述了一项旨在评价案例管理效果的随机实验。该案例管理是由前面的精神病患者项目提供的,与通常的精神健康人员提供的案例相关。这个例子说明了利用实验设计比较服务改革与例行服务的效果。尽管这个设计不能说明相对于无案例管理的对照是否有效果,但说明了一种不同的方法是否比现在的方法更好。这项影响评估的另一个有趣的方面是参加实验的人员样本。尽管合格的案例管理职员代表组是征召来的,但是 25% 的人拒绝参加(当然,这是他们的权利),这样便留下了关于这一实验结果是否能推广到所有合格职员的问题。这是相当典型的服务情形(影响评估中不能覆盖所有合适对象的原因总是多样化的)。即便覆盖所有的合适对象,也可能因为其他原因而不能得到最终的产出测量值。在专栏 8—C 所描述的实验中,评估者是幸运的,96 个参与者中只有 2 个因不能完成服务而缺失,3 个因不能完成一年跟踪而缺失。

专栏 8—B 儿童和青少年心血管健康实验

一项旨在改变在校学生饮食习惯的示范项目实验

按照推荐的饮食容量,一般美国人消费来源于脂肪的能量过多,特别是饱和脂肪,而且,钠的消费量也太高。这些消费类型与冠心病和肥胖症的高发率有关。因此,心、肺、血系统研究所发起了一项旨在改善在校学生营养吸收的随机实地干预实验——儿童和青少年心血管健康实验(The Child and Adolescent Trial for Cardiovascular Health, CATCH)。

儿童和青少年心血管健康实验是一项随机控制实地实验。在这个实验中,基本单位是加州、路易斯安娜州、明尼苏达州和德克萨斯州的 96 所小学。把其中的 56 所学校随机分配作为干预组,另外 40 所学校则作为控制组。干预项目包括举办训练班、给饮食服务人员讲授均衡营养菜谱基本原理以及提供达到这一目的的食谱和菜单。同时,针对教师举办关于营养和实践的训练班,说服管理人员改变学生的体育课程。另外,项目方努力将有关营养的信息送达给参与项目学生的父母。

通过在项目开始时以及 1994 年对儿童进行 24 小时饮食摄入的访谈,结果显示:接受干预学校

的儿童食物总摄入量、总脂肪和饱和脂肪摄入的能量低于对照组学校学生,但是,两者在胆固醇和钠的摄入量上没有差异。因为测量包括24小时内的所有食物,所以,这里的食物类型变化既有学校午餐中的,也有其他膳食中的。还有,在接受干预的学校中,学生血液中的胆固醇水平没有显著下降。重要的是,研究者发现:在被纳入干预组的学校里,学校午餐胆固醇水平并没有下降,也不比对照组学校低。

资料来源:R. V. Luepker, C. L. Perry, S. M. Mckinlay, P. R. Nader, G. S. Parcel, E. J. Stone, L. S. Webber, J. P. Elder, H. A. Feldman, C. C. John son, S. H. Kelder, and M. Wu,"Outcomes of a Field Tiral to Improve Children's Dietary Patterns and Physical Activity: The Child and Adolescent Trial for Cardiovascular Health (CATCH)," *Journal of the American Medical Association*, 1996, 275 (March):768-776.

专栏 8—C　对服务革新增效的评价

费城的一个社区精神健康中心主要为诊断有精神病或有显著治疗史的病人提供强化病例管理。病例管理员采用一种有把握的社区治疗模式(assertive community treatment, ACT)来帮助有各种问题的病人,为他们提供服务,包括住房、康复和社会活动。病例管理队伍由一位病例管理员及其指导下的、受过训练的精神病康复人员组成。

提供精神健康服务的最近趋势是,由自身患有精神病并接受治疗的人提供服务。这个社区精神健康中心感兴趣的是,用服务对象作为病例管理员可能比非服务对象作为管理员更有效果。病人因其自身经历可能对精神病有更深的理解、与病人能建立较好的感情纽带,因此可能产生更合适的服务计划。

为了研究服务对象病例管理相对于精神健康中心通常病例管理的效果,评估者进行了一项实地随机实验。开始时,128 名合格的病人被征召参与这项研究;32 人拒绝参与,剩余的 96 名给予了书面同意,并被随机分配到通常病例管理组或实验组。实验组由精神健康服务对象组成,作为当地服务对象举办的倡导和服务组织的一部分。

分别在分配病例管理的开始、一个月后、一年后,通过访谈和标准化测量搜集数据。测量值包括社会效果(住房、拘留、收入、就业、社会网络)和临床效果(症状、功能水平、住院治疗、急诊室就医、医疗态度和顺从、对治疗的满意程度、生活质量)。为了获得具有统计意义的测量结果,评估者对样本规模和统计分析进行了规划,特别关注了不出现有意义差异的可能性,这将是进行这种比较的重要发现。在 96 个参与者中,94 个在研究期间继续接受服务,91 个被确定在一年的跟踪观察中接受访谈。

除了服务对象病例管理组的病人报告对治疗不太满意和与家人接触较少外,两组的任何其他产出测量值间没有发现统计上的显著差异。虽然有必要对这两个不利的产出进行进一步的研究,但评估者根据主要产出的相似性作出结论:在这一特殊的服务模式中,精神健康服务对象与非服务对象一样,具有同样的能力来管理病例。而且,这种方法会为前面的精神病患者提供相关的就业机会。

资料来源:Phyllis Solomon and Jeffrey Draine,"One-Year Outcomes of a Randomized Trial of Consumer Case Management." *Evaluation and Program Planning*, 1995,18(2):117-127.

专栏 8—D 描述了一项曾经在评估领域做的、与国家政策有关的、最大的和

最好的实地实验。这个实验是为了确定向贫穷、完整家庭(两配偶)提供收入维持费是否会引起他们的有偿工作量减少,即,使得工作积极性受挫。这项研究是5项系列研究中的第一项,每一项与其他项稍有不同,这一系列研究是由经济机会办公室和健康教育福利部(后续机构,健康和人类服务部)管理的,目的是检验不同形式的收入保障及其对穷人和准穷人工作努力程度的影响。5个实验均进行了相当长时间,最长的一个进行了五年多;所有的实验都在参与家庭的合作方面遇到了困难;所有的实验都发现:收入维持费的给付引起了工作积极性的微小受抑,特别是对青少年和有孩子的母亲(次要劳动力)(Mathematica Policy Research, 1983; Robins et al., 1980; Rossi and Lyall, 1976; SRI International, 1983)。

专栏8—D 新泽西—宾夕法尼亚州收入维持实验

在20世纪60年代末期,当关心贫困的联邦官员开始考虑改变福利政策、向所有家庭提供某种年收入保证时,经济机会办公室(OEO)发起了一项大规模的实地实验,以检验项目中的一个关键问题:经济预测理论认为,向贫困家庭给付补充收入会产生抑制效果。

实验开始于1968年,由新泽西州普林斯顿的一家研究公司——计量政策研究公司(*Mathematica*)和威斯康星大学的贫困研究所执行,为期三年。研究的对象群体是收入在贫困线150%以下、男主人的年龄为18~58岁的完整家庭。设定当时的贫困水平和报酬的征税率(根据家庭的收入调节),8个实验条件由各种收入保障组合构成。例如,在其中一种条件下,如果一个家庭没有人挣钱,这个家庭便接受当时贫困水平125%的保障收入。但是如果他们的计划有50%的税,这个家庭的某个成员有收入,他的工作收入每增加1美元,就减少50美分的保障收入。其他的实验条件由30%~70%的税率和相当于贫困线的50%~125%的保障收入水平组成。对照组由没有任何收入的家庭组成。

实验在新泽西州的四个社区和宾夕法尼亚州的一个社区进行。开始时,进行了一项大规模的家庭调查以确定合格的家庭,然后邀请那些合格的家庭参与实验。如果他们同意,随机将这些家庭分配给实验组之一或对照组。实验组家庭报告他们的月收入,如果能转账支付,便将支票邮给他们。在进入项目之前和在实验进行三年中的每个季度末,对参与家庭进行详细访谈调查。访谈搜集的数据包括就业、收入、消费、健康和各种社会心理指标。然后,研究者对这些数据与月收入报告一起进行分析,确定相对于对照组的可比家庭而言,接受报酬是否削弱了他们的工作努力程度(按工作小时数测量)。

尽管开始招收了大约1 300个家庭,但是到实验结束前,22%的家庭停止了合作。其他一些家庭错过了一次或一次以上的调查,或者在不同的时期退出了实验。剩下的连续参与的、可供分析的家庭不到700个。结果是实验组家庭的工作努力程度降低了5%。

资料来源:D. Kershaw and J. Fair, *The New Jersey Income-Maintenance Experiment*, Vol. 1: New York: Academic Press, 1976.

随机实地实验的先决条件

在影响评估中,进行随机评估设计的必要性得到了广泛认同,关于如何提高

随机评估成功率的文献不断增多(Boruch, 1997; Dennis, 1990; Dunfort, 1990)。另外,许多应用于影响评估的实验设计例子,比如本章所引用的例子,证明了随机设计在适当条件满足下的可行性。

尽管随机实验能够最有效地阐明干预效果,但是在影响评估中,随机实验只占相对较小的比例。政治和伦理的考虑会排斥随机化过程,特别是当干预违反伦理或法规而被拒绝时(尽管实验的本意是将某些选择性的干预送达控制组)。即使随机化是可能的和被允许的,随机实地实验设计的实施仍然具有挑战性:如果大规模实施耗费大,需要足够的时间、专门技术、参与者的合作和服务提供者的支持,那么,就只能在条件特别有利的情况下进行实地实验,例如——能够通过抓阄法分配稀有服务时,或者能随机安排有同等吸引力的不同项目时,又或者当项目影响对政策特别重要时。丹尼斯和伯勒奇(Dennis and Boruch, 1989)认为,在进行随机实地实验之前,应该满足五个基本条件(Dennis, 1990):

- 现有实践必须得到改善。
- 预期干预的效力在实地条件下必须是不确定的。
- 没有更简单的替代性干预评估方法。
- 评估的结果对政策必须具有潜在重要性。
- 研究设计必须符合研究者和服务提供者的道德伦理标准。

某些有利于或阻碍随机实验应用于评估影响的条件将在本章后面部分进一步讨论。

近似随机分组

随机选择的可取之处在于随机化本身是一种将合格对象无偏地分配给干预组和控制组的可靠方法;无偏分组要求所有参与者进入干预组和控制组的概率相同。有几种获得干预组和控制组随机选择的替代性方法,在某些恰当条件下也是相对无偏的,因此也在随机选择的可接受范围内。另外,某些案例也说明,尽管群体之间存在差异,那些差异也不会导致测量结果产生偏差。例如,随机选择评估对象的常见的替代方式是,依据某些标准从连续数据中进行系统分组。如果表单不是以某种导致偏差的方法制订的,那么,最常用的随机化替代方法就是由连续表单产生的系统分组(一种经常能实现随机化结果的程序),在将高中学生分配为干预组和控制组时,将所有持单数学号的学生安排在干预组,将所有持双数学号的学生安排在控制组。只要在学生中单双数的安排原本没有区别对待,其结果在统计上将与随机安排一样。当然,如果学校已经给予了女生单数、男生双数,这样,系统偏差就会产生只有一种性别的干预组和控制组。因此,在采用系统选择程序之前,研究者必须确定产生表单的机构是如何完成连续列表的,并判断编码过程是否会在表单的不同部分之间产生不希望的系统差异。

有时,制订的对象表单有微小的难以鉴定的偏差。例如,字母表可能诱使某人将所有姓氏以D开头的人安排到干预组,将以H开头的人安排到控制组。在新英格兰的一个城市,这会导致种族偏见式的选择,因为,许多法国、加拿大人的

名字以 D 开头(如 DeFleur),而很少西班牙人的名字以 H 开头。类似地,如果数字是按顺序排列的,那么数字表可能包括年龄偏差。例如,联邦政府按年龄顺序安排社会保险号,这样,编号小的人一般比编号大的人年长。

在一些情况下,有偏分配是可以"忽略"的(Rog, 1994; Rosenbaum and Rubin, 1983)。例如,在明尼阿波利斯的一项评估,在检验将可放在看护中心的儿童放在家里项目效果的实验中,因为在提名时,咨询机构的能力已到极限,那些不能得到家庭咨询项目服务的儿童便被安排在没有干预的控制组(AuClair and Schwartz, 1986)。"可忽视性"假设是:孩子被送交的时间与孩子对家庭和解的期望无关。因此,如果决定给对象提供或不提供服务(或者,也许是等待批准的服务对象名单列表)的因素与对象的特征无关,那么,结果就处于随机选择的可接受范围之内。

无论是按照无偏分组方法把对象划分为接受项目服务和不接受服务的两个群体,还是虽然配置有偏但能够被忽略掉,都必须对具体情形进行详细审查来帮助慎重判断。如果有任何理由怀疑所研究的事件对具有某些特征的对象的影响较其他单位更多,那么就不是可接受的随机选择,除非能确信无疑地说明那些特征与将要研究的干预和产出无关。例如,饮水中加入少量氟的社区就不能简单地被视为干预组,并拿来与无意评估氟化作用对牙齿健康影响的社区进行对比。因为,采纳加氟措施的社区可能具有与众不同的特点(例如,社区居民平均年龄小,有更具服务导向的管理),这些特点不能被简单地认为与牙齿健康无关,而是代表了此处所指意义上的偏差。

随机实验数据搜集策略

随机实验的两个数据收集策略可以改善项目效果评估。首先,是对产出变量进行多重测量,最好对干预前后的结果都进行测量。在某些情况下,产出变量只能在干预后测量,事先测量是不可能的。除了这种情况,一般的规律是,干预前和干预后对产出变量所做的测量越多,那么所得的净效果的估计值越好。多次历时测量会增加测量的可信度、提供更多的信息以供获得净效果估计值。干预前测量提供的是干预组和控制组实验前状态的估计值,对调整两者之间原来存在的差异以及测量干预的影响是有用的。例如,在一项职业再培训项目中,干预前干预组和控制组的收入测量值会使研究者更好地估计由于培训所获得的收入改善程度,同时提供了一个在结果分析中保持恒定的参照量。

第二个策略是随着干预的展开,进行周期性的数据收集。这些周期性的测量容许评估者建立对干预如何随时间而起作用的描述性解释。例如,如果发现一项为期 6 周的职业再训练项目在前 4 周中产生了大部分效果,这样的评估结果就可能导致在不严重降低项目效果的情况下缩短训练期的建议。同样,多次定期测量能获得对象对服务反应的更全面理解。某些反应可能开始较慢,而后来加快;其他一些反应则是开始强烈,随后很快降到干预前水平。

复杂随机实验

在复杂设计中,影响评估可检验具体干预的内部变化或几个不同的干预。例如,在新泽西—宾夕法尼亚收入维持实验中(专栏8—D),对8种不同的干预进行了测量,每种干预的保障收入量和家庭收入税相互不同。实验中之所以设置不同的干预,是为了考察工作努力程度对包含在不同报酬方案中的工作积极性受挫程度的依赖。关键的评估问题是:报酬对工作的影响是否会随①给付报酬量和②工作收入减少报酬的程度而变化。

顺着这些思路,复杂实验特别适合评估新政策。因为新政策将是或将采用什么形式,事先并不清楚。一定范围的项目变化为可能采纳的具体项目提供了更多的机会,进而增加了影响评估的可概化性。而且,评估的变化能提供信息,有利于指导项目的建构,使效果和效率最优化。

例如,专栏8—E描述了一项在明尼苏达进行的关于福利政策的实地实验。实验条件中涉及两种项目变型,两者都给予被雇佣的福利受益人一定量的救济金。一种有强制性就业和训练活动,一种没有。如果证明这两种项目变型有同样的效果,很清楚,执行没有强制性工作、训练活动和相关管理成本的项目会更节省成本。但是,财政救济金和强制性训练的结合获得了最大的效果。这一信息使政策制定者考虑在较精细但昂贵的项目变型(对收入和就业增加较大影响的)和成本较低但效果较小(却仍然呈积极影响)项目变型之间进行权衡。

专栏8—E 开展福利工作和工作报酬:明尼苏达家庭投资项目

对家有小儿帮助(aid to dependent children, AFDC)项目的一种常见批评是,项目没有鼓励受益者离开福利资助去寻找工作,因为AFDC的报酬明显高于低薪工作所能挣得的收入。明尼苏达州接受了一项来自联邦健康和人类服务部的授权:进行一项实验,鼓励AFDC顾客去寻找工作,如果他们取得成功,允许他们接受比AFDC更高的收入。如果参与者就业,按就业所得,每3美元只减少他们的津贴1美元。明尼苏达家庭投资项目(MFIP)的这些主要变化,使AFDC的救济金增加了20%。还提供了一项儿童看护津贴,以便那些就业者在工作时,孩子能得到照顾。这意味着在这一项目条件下就业的AFDC接受者比他们在AFDC项目中有更高的收入。

在1994—1996年期间,在明尼苏达州的许多县,大约15 000名AFDC接受者被随机安排到三种情况之一:①接受较多的救济金与强制性参与就业和培训活动的MFIP实验组;②只接受较多的救济金不接受强制性就业和培训的MFIP实验组;③继续接受旧的AFDC津贴和服务的对照组。通过管理数据和重复调查,对所有三个组进行督导。产出的测量包括就业、收入和参与教育及培训服务。

涉及第一批9 000名参与者为期18个月项目干预的中期报告表明,实验是成功的。MFIP实验家庭更有可能就业,就业时比对照组家庭有更高的收入。而且,既接受MFIP津贴又接受强制性就业和培训活动的实验组家庭就业比例更高,并且比只接受MFIP津贴的实验组家庭挣得的钱更多。

资料来源:Cynthia Miller, Virginia Knox, Patricia Auspos, Jo Anna Hunter-Manns, and Alan Prenstein, *Making Welfare Work and Work Pay: Implementation and 18Month Impacts of the Minnesota Family Investment Program.* New York: Manpower Demonstration Research Corporation, 1997.

分析随机实验

简单随机实验分析是相当容易的。操作得当的话，可由随机化产生统计上对等的干预组和控制组，所以，对两组结果的比较便提供了对项目效果的估计值。如前所述，统计显著性检验会观察到，如果干预真的没有效果，那么误差值就会大于效果值。专栏 8—F 列举了在简单随机实验基础上进行分析的例子。首先，简单比较实验组和对照组；进而，对结果进行分析；然后，借助于更复杂的多元回归模型进行分析。

正如可预期的那样，复杂随机实验要求相应的复杂分析模型。尽管简单的方差分析足以获得总体效果的估计值，但更为精确的分析技术对揭露内情和解释结果的作用更大。例如，严谨的多元分析可以提供更为精确的干预评估，并允许评估者提出简单随机实验无法回答的问题。专栏 8—G 描述了一项复杂随机实验如何运用方差分析和因果模型来进行分析的实例。

专栏 8—F　对随机实验的分析：巴尔的摩生活项目

巴尔的摩生活实验（LIFE）项目的设置是为了评估给予刑满释放者少量财政资助是否有助于他们过渡到正常生活、减少他们被捕和重新入狱的可能性。由于大多数囚犯在监禁期不能积累工作存款，他们就不能获得失业保险金。所以，项目仿照失业保险金设定财政资助。

将从马里兰州监狱释放、返回巴尔的摩的人员随机安排到干预组或控制组。告诉实验组成员，只要他们失业了就够资格领取 13 周、每周 60 美元的报酬。同时，告诉控制组成员，他们正在参加一项研究项目，但不给报酬。研究者定期采访这些参与者，并且对他们的拘留记录进行为期一年的督导（从每位参与者的获释时间开始算起）。释放后一年间的拘留记录结果见表 8—F1。

表 8—F1　释放后第一年的拘留率

拘留指控	实验组 (n = 216)	对照组 (n = 216)	差异
盗窃（如，抢掠，入室行窃，盗窃）	22.2%	30.6%	-8.4
其他严重犯罪（如，谋杀，强奸，袭击）	19.4%	16.2%	+3.2
轻罪（如，破坏治安，公开酗酒）	7.9%	10.2%	-2.3

表中的结果被认为是主效应，构成了实验结果的最简单表达。因为随机化使得干预组和控制组除了干预之外，在统计上是对等的，所以，可以设想它们之间拘留率的差异是由干预和任何随机变异引起。

实验的实质性结果表现为表右边的最后一列，表示干预组和控制组之间各类犯罪拘留率的差异。在释放后一年里，-8.4% 的盗窃罪差异表明了期望的潜在干预效应。那么，问题便成了在给定样本大小（*n*）情况下，8.4 是否是在期望的误差范围内。在这种情况下，可以应用各种统计检验方法，包括卡方检验、T 检验和方差分析。因为两组间差异的方向是由干预的期望效果决定的，所以研究者使用了单尾 T 检验。结果表明：-8.4% 或者更大的差异在同样大小样本的 100 次实验中发生的次数小于 5 次（统计显著水平为 $p \leq 0.05$）。于是，研究者得出结论，至少对盗窃而言，这一差异是显著的，可以认为干预获得了期望的效果。

其他类型的犯罪没有表现出足够的、达到T检验标准的差异。换句话说,按照常用的统计标准($p>0.05$),干预组和控制组之间的差异仍在随机误差足够解释的范围内。

如果是这样,下一个问题就是:在政策意义上,这些差异是否足够大?盗窃犯罪减少8.4%能证明所付报酬和相应的管理开支是合理的吗?为了回答后面一个问题,劳动部进行了成本—收益分析(参见第11章的讨论),表明收益远远大于投资。

用多元回归分析盗窃犯罪数据的更复杂和更大程度上利用信息的方法列于表8—F2之中。提出的问题与前面分析中的问题是一样,但是,多元回归模型考虑的事实是:除了给付的报酬外,其他许多因素也可能影响拘留。多元回归分析在比较对照组和实验组拘留比例的同时,在统计上控制了其他因素的影响。

实际上,对干预组和控制组之间的比较包括了分析中所用的、其他变量的每个层面。例如,巴尔的摩失业率在实验两年期间的变化是:某些囚犯在容易找到工作的时候被释放,而其他人被释放的幸运次数较少。把释放时的失业率添加到分析中,就减少了因这一因素引起的个体间差异,因而净化了干预效果的估计值。

表8—F2　因盗窃罪拘留的多元回归分析

自变量	回归系数(b)	回归系数b的标准误
干预组成员	-.083*	.041
释放时的失业率	.041*	.022
释放后当季度工作的周数	-.006	.005
释放时的年龄	-.009*	.004
第一次拘留的年龄	-.010*	.006
以前的盗窃拘留	.028*	.008
种族	.056	.064
教育	-.025	.022
以前的工作经验	-.009	.008
婚姻	-.074	.065
假释	-.025	.051
截距	.263	.185

$R^2=.094*$;$N=432$;*表示$p\leqslant.05$的显著性水平

注意,增加到表8—F2多元回归分析中的所有变量,是从以前的研究中得知的影响刑满释放人员或就业机会的那些变量。这些变量的增加在相当程度上增强了实验结果的可信度。每一个系数表示释放后拘留概率的变化与每个自变量相关。这样,与实验组的相关系数-.083意味着干预使盗窃罪拘留率减少了8.3%。这与表8—F1中显示的结果是密切对应的。但是,因为对分析中其他变量的统计控制,回归系数的误差期望值降低很多,直到在100次实验中只有2次。所以,多元回归结果提供了干预净效果的更精确估计值,也告诉我们释放时的失业率、释放时的年龄和第一次拘留时的年龄,以及以前的盗窃拘留情况,都是对拘留率有显著影响的因素,所以,这些因素影响到项目的产出。

资料来源:P. H. Rossi, R. A. Berk, and K. J. Lenihan, *Money, Work and Crime: Some Experimental Evidence*. New York: Academic Press, 1980.

专栏8—G 复杂随机实验的分析:TARP研究

专栏8—F中描述的巴尔的摩生活实验令人鼓舞,劳工部因此进行了一项更大规模的实验——利用两个州现存的机构管理发放给前重罪犯失业保险金。这个新项目的目的也是一样的:使前重罪犯够资格领取失业保险金,进而减少他们靠犯罪来获得收入的需要。但是,新项目叫做刑满释放过渡帮助(TARP),其中有不同的干预设置,包括合格接受津贴的不同期限和对每1美元就业所得减少津贴(税率)的不同比例。

干预的主要效果见表8—G中的方差分析(为了简便起见,仅将德克萨斯州TARP实验的结果列出)。干预对与财产有关的拘留没有影响:干预组和控制组的差异未超过抽样误差。但是,干预对释放期间的工作周数影响很大:接受报酬的前重罪犯平均工作时间少于对照组人员,而且,其差异达到了统计上的显著水平。简而言之,增加报酬似乎没有构成对犯罪的抑制,但却成功地抑制了就业!

表8—G 与财产有关拘捕的方差分析(德克萨斯州数据)

A. 释放期间与财产有关的拘留

干预组	拘留平均数	被拘留百分数	n
给付26周报酬,100%税	.27	22.3	176
给付13周报酬,25%税	.43	27.5	200
给付13周报酬,100%税	.30	23.5	200
不付报酬,安排工作[a]	.30	20.0	200
被访谈的对照组	.33	22.0	200
没被访谈的对照组[b]	.33	23.2	1 000
方差分析的 *F* 值	1.15(p=.33)	.70(p=.63)	

B. 释放期间工作周数

干预组	工作周平均数	n
支付26周报酬,100%税	20.8	169
支付13周报酬,25%税	24.6	181
支付13周报酬,100%税	27.1	191
不付报酬,安排工作	29.3	197
被访谈的对照组	28.3	189
方差分析的 *F* 值	6.98(p<.000 1)	

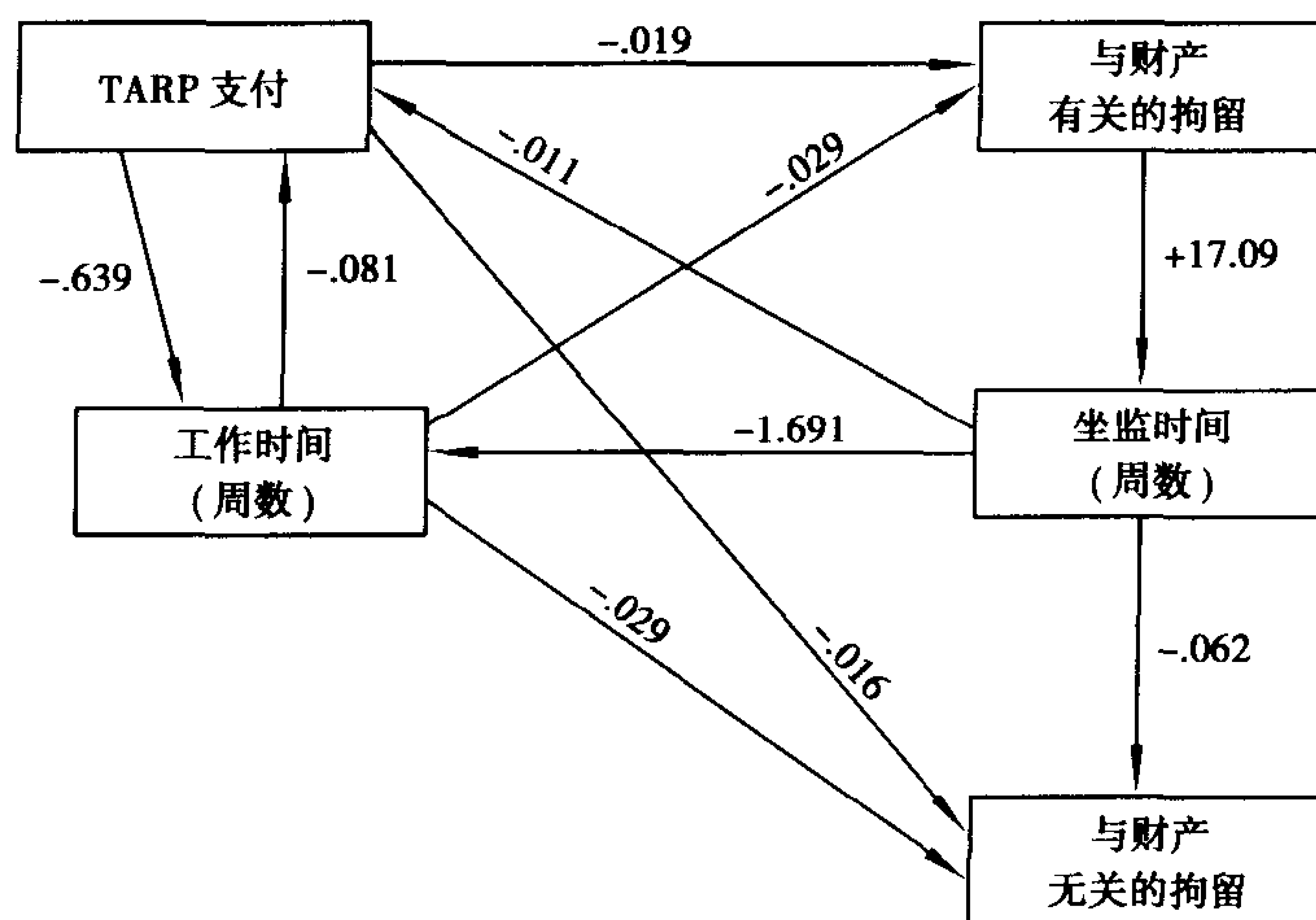
a. 提供给这一干预组中的前罪犯特别的工作安置服务(很少被采纳),根据工作需要提供工作工具和制服。但支付给这些前罪犯的报酬很少。

b. 只是通过拘留记录对对照组进行观察,所以,没有关于工作周数的资料。

总之,以上结果似乎表明实验干预没有以预期的方式起作用;而且实际上,还产生了预期不到

的效果。但是,这种方差分析只是分析的开始。对评估者来说,这些结果暗示着一些平衡过程可能在其中起作用。从犯罪学文献可知,前重罪犯的失业与他们再次被捕入狱的可能性增加有关。所以,研究者推断失业津贴挫伤了工作积极性,具体表现在接受较多工作津贴或较低税率的参与者的工作周数较少,进而使犯罪行为增加。另一方面,报酬也会减少靠从事犯罪而挣得收入的需要。这样,报酬在减少犯罪中的正效果可能会被报酬给付期间减少就业的副效果所抵消,以致对拘留的实际总效果为零。

为了检测对实验结果的这种"平衡效果"解释的合理性,研究者建构了一个因子模型,如图 8—G。在这个模型中,负系数是报酬对就业的影响(工作积极性受挫)和对拘留的影响(期望的干预效果)。反过来,失业的平衡效果表现为拘留和就业之间的一种负系数,这表明工作周数减少与拘留增多相关。图 8—G 中显示的系数来自于结构方程模型分析。如图所示,在德克萨斯和佐治亚州的数据分析中均呈现出了被假设的关系。



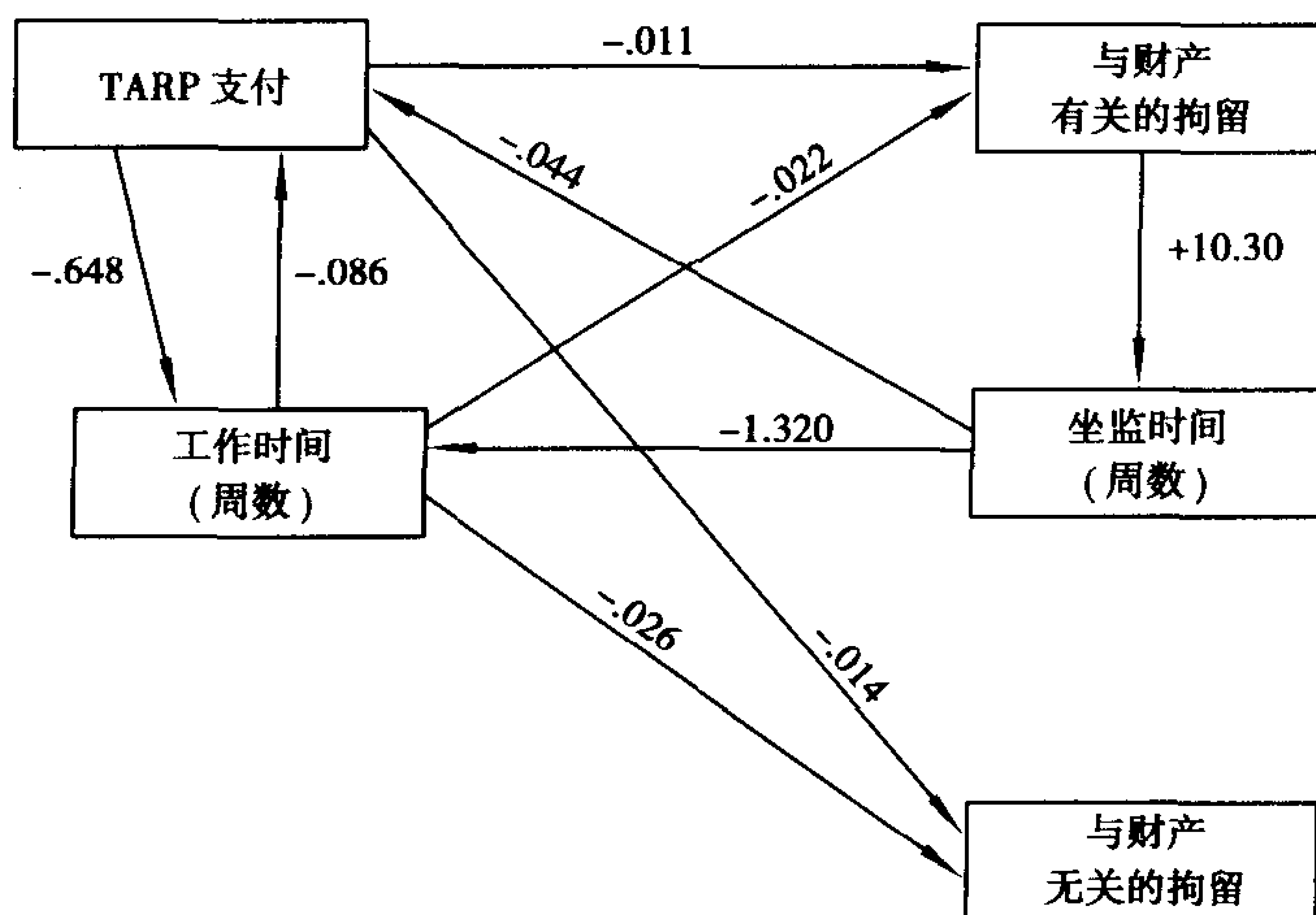


图 8—G

因此,与复杂多元分析相结合的复杂实验表明,干预的净效果是可以忽略不计的。即使如此,

也提供了对评估结果的某些解释。特别是有关数据证明了:增加报酬所期望的作用是减少犯罪行为,但是,一个成功的项目不得不设法弥补随之而来的副效果(工作积极性受挫)。

资料来源:P. H. Rossi, R. A. Berk, and K. J. Lenihan, *Money, Work and Crime: Some Experimental Evidence*. New York: Academic Press, 1980.

随机实验的局限性

随机设计原本是为实验室和农业田间研究而产生的。尽管其自身的逻辑思想很适合评价社会项目的影响,但却并不适用于所有的项目情形。在本节,我们将评述其局限性。

处于早期实施阶段的项目

正如本章已经列举的某些例子那样,示范性项目的随机实验能为政策和项目设计提供非常有用的信息。但是,一旦项目设计被采纳并进入实施,在项目稳定和操作成熟之前,并不能马上知道随机实验影响评估是否良好。在项目实施的早期阶段,为了完善干预及其他服务的送达,经常需要改变项目的各种特征。尽管随机实验能将项目的产出与未干预对象的产出进行比较,但是,如果项目在实验过程中发生了变化,其产出将不能提供应有的信息。如果在测量参与者的产出之前,项目发生了略微的变化,不同类型干预的效果混在一起,也就无法确定任意给定干预的效果。所以,昂贵的实地实验最适合对实验过程中干预活动始终如一的项目效果进行评估。

伦理的考虑

应用随机实验经常遇到的障碍是,某些管理者对随机化有伦理上的疑虑,认为是对控制组实际利益的专横和任意剥夺。这种批评的理由是:如果值得对一个项目进行实验(例如,如果项目可能有助于研究对象),那么从那些需要项目服务的对象撤销潜在的有帮助的服务,便是一种实实在在的伤害。因此,这样做是不道德的。相反的论点也显而易见:人们通常不知道干预活动是否有效,实际上这也是实验的理由。因为研究者事先不知道一项干预是否会有帮助,所以实验不是在剥夺控制组的某些被认为将是有益的东西。

有时,一项干预可能意味着某些危害,在这种情况下,决策者仍然可能会勉强批准随机化。例如在定价实验中,某些实验组家庭的非食物日常费用会增加。研究者则认为,他们对实验家庭许诺,任何这样的超支在研究结束后都会得到补偿。当然,这种偿还的许诺改变了干预的特性,可能会助长对家庭用具的过量使用。

一般来说,最引人注目的伦理性反对意见与控制组的状态相关。如果常规服务对问题是有效的,那么为了评估一项常规服务的替代性服务效果而撤销原来的

服务就是不道德的。例如,我们不会剥夺学龄儿童的数学教育以使他们能组成一个评估新数学课程的控制组。但是,在这些情况下,重要的问题不是引入新课程是否比没有相应教育更好,而是是否比现在的教育更好。因此,合适的实验是,在新课程与现有课程之间进行比较,而不是让学生去上没有任何保证的课程。

当项目资源不充足和不符合需求时,实验控制条件下的随机分组会出现伦理上进退两难的困境。一方面,这个过程会导致随机选择那些不太合格的对象接受项目服务。另一方面,如果不能将干预送达所有合格的对象,随机化本身就会变成一种谁将得到公平对待的决策,进而引发争议,因为所有对象本应有同等的被择机会。假若干预的效果很不确定,这种做法也许能被接受。问题是,当服务提供者相信干预有效(尽管他们经常缺乏实验证据),他们可能强烈反对通过抽签安排服务,而坚持让最需要的对象有优先权。正如将在下一章讨论的一样,这种情况很适合回归—间断点设计,尽管对随机设计来说,回归—间断点设计有很大的问题。

实验和实际干预送达之间的差异

评估中,随机实验的另一个局限是:在实验类型的影响评估中,干预的送达很可能与在项目实施时的实际送达在关键方式和具体情形上有所不同。因为只有有限的方法能使干预送达,所以利用这些标准的、容易送达的干预,研究者相对能确信实验干预类似于一个完全得到实施的项目干预。然而,劳动力较密集、技能较高的干预(譬如工作安置服务、咨询、教学等)在实地实验中可能比项目实施时以更高的、对设计者意图更了解的方式送达。实际上,如在第 6 章所见,项目督导的目的之一就是督导干预在实施过程中发生变化的可能性。

由于存在这样的可能性,所以需要至少两轮实验:在第一轮实验中以最单纯的形式对干预进行测量,在第二轮实验中对通过公共机构的有效服务送达进行测量和比较。专栏 8—F 和 8—G 描述的给刑满释放人员提供就业保险津贴的劳工部项目,两个实验阶段均采用了这种策略。第一阶段工作由在巴尔的摩进行的涉及 432 个从马里兰州监狱释放囚犯的小规模实验组成。研究者在释放前对囚犯进行选择,提供给他们津贴,并对他们的工作和拘留情况进行为期一年的观察。结果如专栏 8—F 显示:接受失业保险津贴的实验组在释放后的一段时间里行窃拘留减少了。

第二轮更大的实验是在乔治亚州和德克萨斯州进行的。每个州有 2 000 名刑满释放人员(专栏 8—G)。在这个实验中,津贴是通过各个州的职业保障机构来管理的,并且由他们和州监狱系统联合在释放期间跟踪调查项目对象。如果该项目当时得到联邦立法通过,第二阶段的实验管理系统就接近于实际实施的情形。但是,第二阶段的结果表明:当按既有就业保险机构(Employment Security Agency)的规则和程序进行管理时,给付报酬是无效的。

时间和成本

应用随机实地实验时,一个突出的障碍就是耗资耗时,大规模的多点实验尤

其如此。因此,不应该用随机实地实验来评估决策者很可能不会采纳的项目设想;或者,当没有重要的相关利益方对项目影响的证据感兴趣时,也不适合运用随机实地实验方法来评估业已实施的项目。另外,当急需产出信息时,也不应采用这类实验。为了强调最后一点,应该注意到新泽西—宾夕法尼亚州收入维持实验(专栏8—D)耗资34 000 000(按1968年的美元计)美元,从设计到发表结果花了7年多的时间。西雅图和丹佛的收入维持实验花的时间则更长,在收入维持作为一项政策从国家议事日程消失后很久,最终的实验结果才面世(Mathematica Policy Research, 1983; Office of Income Security, 1983; SRI International, 1983)。

实验的完整性

最后,我们应该注意到,一项随机实验的完整性很容易受到威胁。尽管在评估的开始阶段,随机形成的干预组和控制组在统计上是对等的,但随着实验的进展,一些非随机过程可能威胁到对等性问题。不同损耗率可能引起干预组和控制组之间的差异。例如,在收入维持实验中,接受较少津贴的干预组及控制组家庭更可能停止参与实验。我们没有理由相信,在只有很少成员退出的条件下,能确保其他条件的完全对等,在项目效果评估中,干预组和控制组的可比性是与相应潜在偏差相妥协的产物。

另外,送达一项"纯项目"是困难的。尽管评估者可能设计一项实验去评估某种干预的效果,但对实验对象所做的每件事情都会成为干预的一部分。例如,TARP实验(专栏8—G)原本是为了测量刑满释放后财力资助的效果,但是这项资助是通过现有的州机构管理的,后者的操作过程随之变成了干预的一部分。事实上,不受这种过程影响的大规模随机社会实验几乎不存在。当然,即便随机化被折衷到某种程度,随机实验的结果如果分析得当,仍然可能在可信度上优于下一章将要讨论的非随机设计。

小　结

- 影响评估的目的是判断项目干预对预期产出的影响。随机实验是评估的旗舰,如果操作良好,可以提供关于社会项目效果的最可信结论。
- 影响评估可能在项目操作过程的多个阶段进行。但因为严格的影响评估都需要关键性资源的支持,所以评估者应该考虑项目影响评估的进行是否能得到环境条件的许可。
- 所有影响评估研究设计的方法论基础都建立在随机实验逻辑之上。这一逻辑的主要特征是研究中的目标对象(干预组和控制组)是随机分派的。在准实验设计中,分组任务是通过非随机方式完成的。评估者必须判定构成"足够好的"研究设计的每一个条件和要求。
- 随机实验的主要优势在于:通过保证干预组和控制组统计上的对等性来对项目干预进行评估,从而分离出项目干预的效果。除了接受的项目干预之外,严格意义上的对等群体在成员构成、观察期间的经历以及对所研究项目的倾向上都是一致的。但实际上,作为集合体的群体在与产出相关的一些特征方面具有一致性就已经足够了。

- 尽管偶然性在随机选择的任何两组之间会产生差异,但统计显著性检验可以让研究者把由偶然性产生的误差从由项目干预所产生的效果中分离出来。
- 在影响评估中,分析单位的选择是由干预和干预所指向的对象特性所决定的。
- 通过一些程序或情境设定,能够产生可接受的近似随机选择,如根据项目吸纳对象的能力,在给定时间内,每隔一个间隔将相应个体选入名单或列入服务对象。但是,只有当这些替代方法产生的干预组和控制组在预期干预或产出方面具有相同特征时,才可以充分替代随机化过程。
- 尽管在影响评估中干预后的产出测量值是至关重要的,但干预前和干预期间的测量以及以后的重复测量,提高了测量的信度和效果估计值的准确性,使研究者能够检测干预随时间所产生的效果变化。
- 复杂影响评估可以核查几个干预或单个干预不同阶段的变化。这些实验分析要求更为复杂的统计技术。
- 尽管随机实验很严密,但对某些影响评估来说仍可能是不合适或不可行的。当运用于项目实施的早期阶段时,或干预以实验不易捕捉的方式变化时,随机实验的结果可能并不明朗。另外,如果项目各方认为拒绝给控制组提供项目服务是不公平或不道德的,也可能不允许进行随机选择。
- 实地实验需要广泛的资源、专门的技术、研究积累、时间保障和项目对正常服务送达中断的容忍力。当然,也可以创造出某些人为的情境,使得此类情境下的项目服务送达不同于实践中正式送达的干预。

基本概念

控制组(Control group):一组没有接受项目干预的对象,用于与一组或多组接受项目干预的对象进行产出比较。可以参照干预组来理解。

干预组(Intervention group):接受项目干预且产出观测值要与一组或多组控制组的观测值进行比较的一组对象。可以参照控制组来理解。

准实验(Quasi-experiment):通过非随机分配程序构成干预组和控制组的影响研究设计。

随机选择(Randomization):按照概率原理将项目对象安排到实验组和对照组的过程。

随机实地实验(Randomized field experiment):是基于特定项目背景而进行的一种影响研究设计。在这种研究设计中,随机分派干预组和控制组成员,通过比较两组对象的测量结果来判断项目干预效果。可以参照控制组、干预组来理解。

分析单位(Units of analysis):在项目影响评估中,进行项目产出测量的单元;相应地,所得测量数据在分析单位的基础上可供评估者进行处理与分析。分析单位可以是个人,也可以是家庭、邻里、社区、组织、行政区划或地理区域等任何此类实体。

项目影响评估——备选设计 9

由于可以对项目产出进行无偏估计，影响评估更容易选择随机实验方法。不过，更为常见的是，评估者必须依赖非随机设计。在本章中，我们将讨论项目产出评估中干预组和控制组的非随机设计方法。这样的设计方法通常用于不能将参与和不参与项目的对象进行随机分组的情况中。我们也将讨论另一种利用反身控制的设计（对象与其自身相比较）。无论如何，在这些备选设计中，没有哪一种能达到随机实验估计项目产出的无偏程度。

我们在第 8 章已经讨论过,随机实地实验是最科学的、可信的影响评估设计方法。如果进行得好,对项目影响的估计将是无偏的,也就是说,没有内在的缺陷去高估或低估项目产出。评估者进行影响评估的目标,当然应该是对项目的真实效果做出公正、精确的估计。这就是在可行的条件下,人们通常选择随机实验方案的原因。

不过,当随机设计不可行的时候,评估者也可以有其他选择。但是,这些备选设计有一个共同的问题:即使执行得很好,对项目产出的估计可能还有偏差。这些偏差会系统地增大或者缩小项目效果,而且,作用的方向通常不可预测。这样的偏差当然会影响到项目各方的利益。如果这些偏差使得一个无效或者有害的项目呈现出有效的状态,其对项目参与者将是不利的。而投资者和政策制定者担心的是在这种情况下,浪费资源且于事无补。另外一方面,偏差也可以使一个真正有效的项目看起来无效或有害,不公平地降低对项目人员工作成果的认可,从而很可能让赞助者减少或者取消投资。

对于评估者来说,在进行影响评估设计时,最关心的问题是如何使估计偏差最小化。相应地,进行非随机设计时,主要关注如何使偏差的可能性最小。因为不能确保没有偏差,所以,非随机设计的影响评估在这方面或多或少存在问题。为了了解为什么以及研究者应当如何应对,我们必须知道偏差从何而来。我们将了解非随机设计的各种形式,以及当随机设计不可行时如何运用非随机设计来估计项目产出。

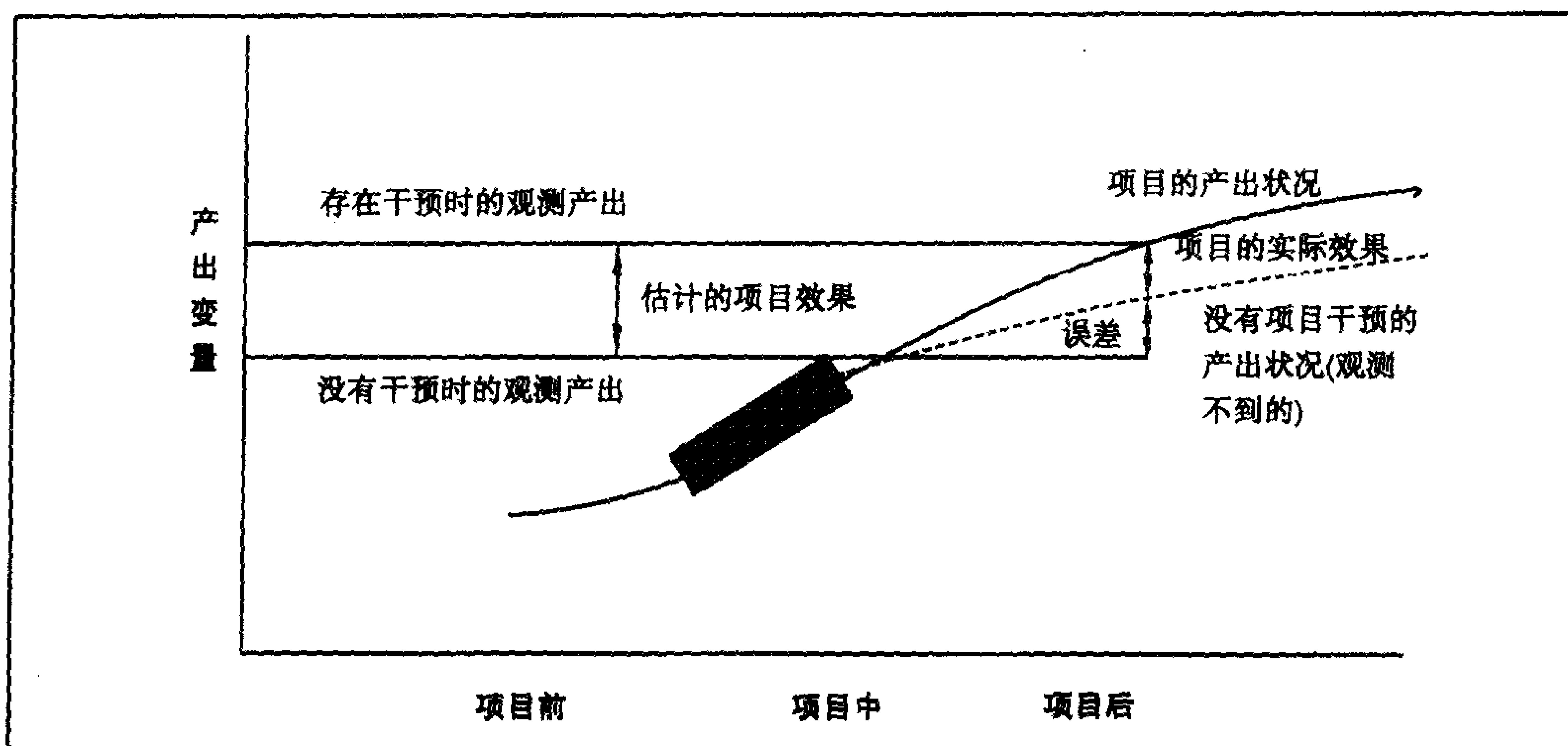
项目效果估计的偏差

在项目目标确定之后,评估者就能够观测产出状况了,并且通过效度和信度的测量,以可接受的精确度来描述项目产出。第 7 章已经谈到,项目效果是干预组和控制组(未接受项目干预)在产出测量上的差值。当然,这种比较被限定在其他条件对等的前提下。偏差来自于对项目产出的测量值或没有项目介入时的测量值的高估或低估。在评估数据中,当上述两者任意一个产出被误判时,对项目效果的估计就会比实际效果大或小,这就是估计偏差。

第一个问题是,在干预进入后,针对目标人群测量可观测产出时所出现的偏差。通过采用真实的、贴切的、全面的测量手段,来测量在目标人群中所产生的项目产出,这一类偏差是比较容易避免的。影响评估中的偏差经常来自于研究设计系统性地高估或低估没有项目干预情况下可观测的产出。关于这类偏差,可以参见专栏 9—A,这个专栏使用了第 7 章(专栏 7—A)中展示项目效果的图。由于无法在施加干预的项目中观测到不施加干预的效果,所以,我们没有现成的手段来测定项目效果估计中何时会出现这类偏差。这种内在的不确定性,使得估计偏差很可能对影响评估发生潜在威胁。在这里,有必要提及随机设计的益处,尽管这种设计在某些案例中可能会错误地估计没有干预时的产出,但随机化

方法使得这种情况随机发生，高估和低估的概率相近，而且不存在系统偏差。

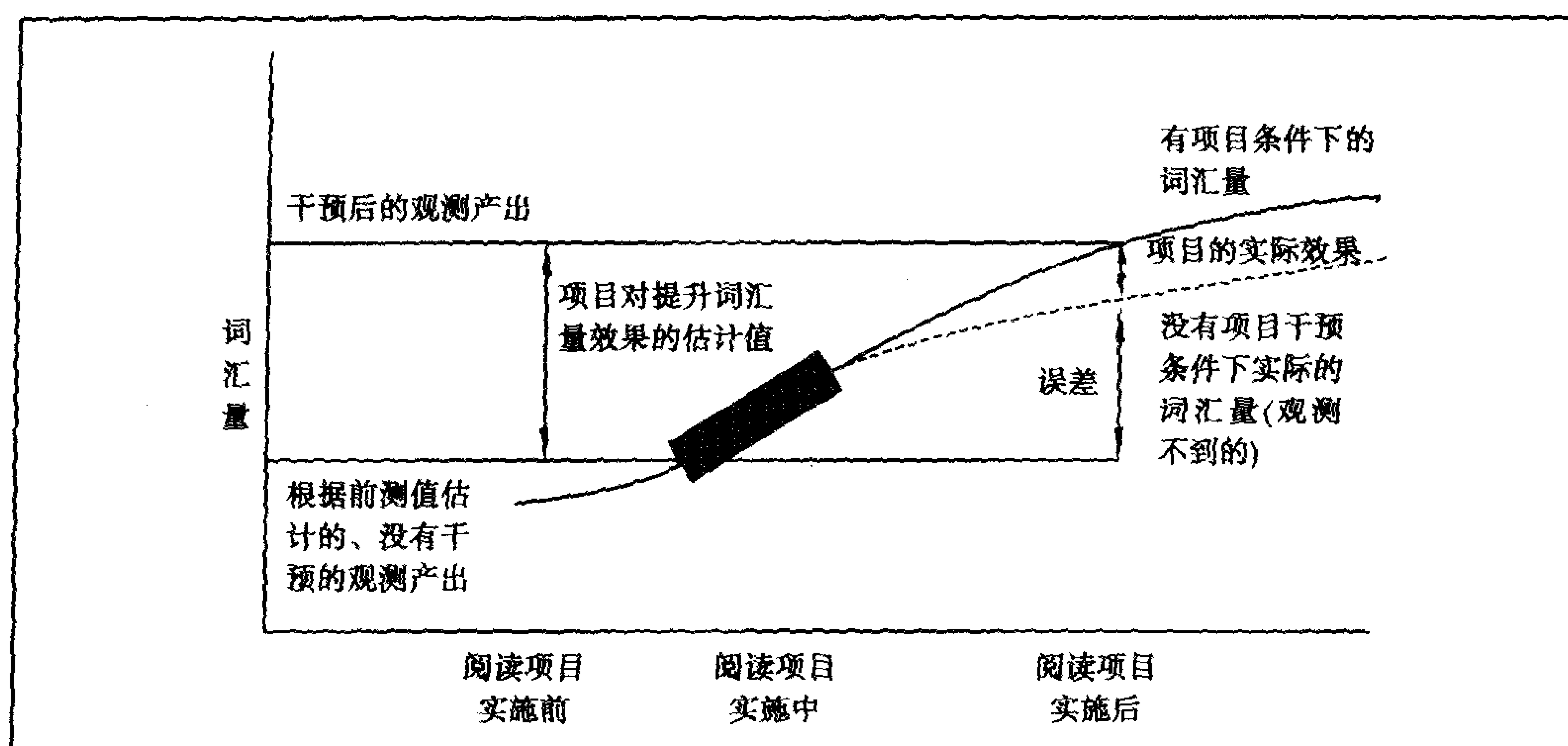
专栏9—A　项目效果评估中的偏差示意图



举个例子来说明有可能存在相当明显的偏差。假设正在评估一项旨在提高儿童词汇量的阅读计划，并且我们有合适的词汇测量工具。我们用这个词汇测量工具在项目开展的前后，分别测量儿童的词汇量。由于这个测验对于项目着眼的词汇量提升具有信度和效度，并且非常敏感，所以，我们对参加项目后孩子们的词汇状况有信心做出真实的描述。为了估计项目对词汇量增加的效用，我们应该将参加项目后孩子们的词汇水平与假定没有参加项目可能具有的水平进行比较。为了达到这个目的，我们决定使用在项目开展前词汇测量的结果。由此可以假定，如果没有这个项目，孩子们的词汇水平将不会发生改变。随后，我们将项目实施后得到的测量均值减去前测的词汇量均值，就得到项目效果的估测值。这种处置流程图见专栏9—B。

不过，在对上述项目效果的估计中存在着偏差，因为孩子们的词汇量事实上并不是静态的，而是随着时间不断增长的（并且在通常情况下根本不可能减少）。这就意味着，即使没有上述项目的存在，孩子们的词汇量也会增长，尽管增长的幅度不一定高于有项目帮助时的情况。实际上，孩子们自发增长的词汇数量也被计入到对项目实际效果的估计之中，这就是我们估计过程中偏差的来源（专栏9—B展现了这一点）。由于在人类行为的许多方面都存在着这种自然的变化，所以前一后测总是会在对项目绩效的估计中产生偏差。对于从事影响评估的工作者而言，很不幸的是，并不是各种形式的偏差都像上述例子中一样明显地威胁到项目的影响评估，并且能被轻易地发现，下面我们就要对此展开讨论。

专栏9—B　前后变化测量基础上 提升儿童词汇量的阅读项目效果评估中的偏差



选择性偏差

在前面的章节中已经说到,影响评估设计的最常见形式是包含两组个体或者组织的比较设计,其中项目组接受项目干预,控制组则不接受干预。因此,对项目效果的估计基于合适测量手段(例如,项目组的平均得分减去控制组的平均得分)之下项目组与控制组之间的测量差别。在都不接受干预的情况下,只有两组的平均得分在实验时间间隔内始终不变,上述的简单差别才是对项目效果的无偏估计。如果上述预设成立,那么两组之间任何数值上的差别都是项目的效果。将个体随机分配到干预组和控制组,就可以认为两组是对等的,并通过统计显著性检验来判断差异的显著性。

如果分组不是随机指定的,那么,上述等值预设能否成立就值得怀疑。无论各组看起来如何等值,只要不是随机分组的比较设计都被称为**非对等比较设计**(Nonequivalent comparisom design)。这一称谓强调在没有项目干预的情况下,不能作出结果对等的预设。

当不能作出对等预设时,就无法避免原有组间差异对项目效果估测产生偏差效应,这就是所谓的**选择性偏差**(Selection bias)。对于使用非对等(即非随机化的)分组比较设计的任何影响评估,选择性偏差都是真实地估测项目效果的内在威胁。

选择性偏差的得名,是因为与随机分组相对应,我们完全不知道分组的哪些过程会产生影响。例如,对自愿参与的个体实施项目干预,而将非自愿的个体划入到控制组。自愿参与的个体已经自主选择了将参与哪个群体。选择性偏差是在没有项目干预条件下,任何自愿者和非自愿者之间的产出测量差别。由于不

太可能知道自愿者与非自愿者之间的所有相关差别,在确定这类偏差的性质和范围时所能做到的就相当有限。

选择性偏差并非只指类似自愿者和非自愿者这样由非随机分组造成的偏差。选择性偏差常常表现为更加隐蔽的形式。例如,假设评估校园毒品预防项目影响的评估者,发现另一个临近的学校没有参与这个项目,且在其他方面很类似。评估者可能就会拿临近学校的学生作为控制组,来同参与了项目的本校学生进行对比,比较两个群体在学年末毒品使用的状况。即使在学年初,两校学生的毒品使用状况类似,然而,研究者怎么能够确定在排除项目影响的条件下两校学生的毒品使用状况在学年末仍然会一样呢?在孩子们居住的地区和他们所上的学校,有许多个人的、文化的和经济的因素对他们行为发生着影响。这些因素发挥着某种"选择"作用,使得某些孩子上这所学校,而另一些孩子上另外一所学校。那些影响孩子们上某所学校的因素,也可能同样影响到他们在学年内的毒品使用情况。一旦存在这种状况,对两所学校毒品使用情况和毒品预防项目效果的任何估测都可能存在选择性偏差。

导致项目组或控制组成员产出数据损失的自然或人为过程,也是引发选择性偏差的一个原因,这种情况被称作**损耗**(Attrition)。损耗可能产生于下面两种情况:①对象从项目组或者控制组离开,并且再也接触不到;②评估对象拒绝就项目产出进行测量。问题的关键是在项目结束后的测量中出现了样本缺失,原来被分配到控制组或项目组的某些个体被剔除。无论何时,常用的精确随机过程(例如使用随机数表或者抛硬币法)以外的任何因素,都会带来不同程度的损耗。也就是说,项目组中缺失的产出数据,不能假设与控制组中缺失的产出数据具有同样的特征。由于选择性偏差的存在,排除损耗之后保留在各组中的数据,其可比性也会产生变化。

随机分组设计显然也不能摆脱因损耗所导致的选择性偏差。随机分组在分组之初是统计均衡的,但只有在项目结束之后的测量中也保持均衡,才能减少选择性偏差。因此,选择性偏差由于控制组及项目组中存在的损耗而产生,会造成产出数据不能涵盖最初随机选定的每一个测量单元。因此,为了保证随机实地实验的效度,研究者就应该防止或降低产出测量上的损耗,不然,这一实验设计就极可能蜕变为非对等比较分组实验。

应该注意到,危害研究设计的这类损耗是由产出测量中的样本丢失带来的。如果干预对象从项目中退出以后仍然参与产出测量,这就不构成选择性偏差。如果干预对象没有完成项目,这会影响到项目执行,但是不管执行结果怎样,都不会危害到对项目影响的研究设计。因此,研究者应该尽量获取每一个项目组成员的产出测量值,无论他们是不是完整地参与了整个项目过程。类似地,控制组每一个成员的产出数据也应该尽量获取,无论他们是不是提前退出了项目或者接受了其他相关服务。如果得到了完整的项目产出数据,那么,比较两组测量值的效度就能够得到保证。项目组接受的干预或者控制组接受的类似服务如果不完全,那么,对项目效果估计以及比较的显著性将受到影响。无论目标群体的

最终项目效果如何,都会小于完整执行项目情况下的产出。如果控制组接受了服务,那么,对项目效果的估测就代表了相对于控制组的最后收获,无论项目组的执行多么完善。

总之,选择性偏差并不只是适用于对目标群体的最初分组,也关涉到在产出测量阶段各组数据的可用性。也就是说,选择性偏差包含了诸多情形,在这些情形中,撇开那些直接关联项目参与的特征不谈,项目组和控制组产出测量比较的某些内在特征差异,也会影响到产出测量的状况。

偏差的其他来源

除了选择性偏差之外,还有一些因素也会导致影响评估结果的偏差,这些因素总体上与干预阶段中接受项目影响之外的其他事件和经验相关。通过对项目组和控制组进行比较,从而估测项目效果,这不仅要求两组样本在相关于产出测量的特征上对等,也要求除了是否参与项目之外,研究中的其他经验也对等。如果某一项目组具有另一组没有的其他经历,并且这些经历会影响到项目产出,那么两个项目组结果的差别就不仅反映项目的效果,也反映其他经历的效果。其他因素带来的效果当然会使项目的效果放大或缩小,构成评估的偏差。

即使是在前—后测设计中,当接受项目干预的个体在与接受干预之前对比时,干扰事件或干扰性经历仍然是潜在的问题。有时候,个体的经历或是在项目进行过程中发生的重大事件,即便与项目本身无关,也可能影响到项目产出。这样,前后比较就会产生出对项目效果的有偏估计。譬如,对于儿童来说,在阅读项目进行中,词汇量的自然增长就是一个干扰因素(专栏 9—B)。

对评估者来说,困难在于项目的运作需要社会环境的支持。在这个环境中,平常的或自然的事件演进不可避免地影响到项目关注的结果。例如,许多人能从急症中自然地恢复,因为情绪的放松显然有助于这类疾病的康复。因此,检验治疗某种疾病——比如说感冒——方案的医学试验,必须区分治疗的效果与没有治疗情况下的病情变化;否则,对治疗方案效果的估计就具有很大的偏差。社会干预也具有类似性。训练年轻人掌握某项技能的项目必须认识到,有些人即便不参加项目也能够掌握这些技能。类似地,对于评估减轻贫困的某些项目来说,必须考虑到某些家庭和个人在没有外援的情况下,经济状况也会逐渐转好。

在对项目效果的评估中,那些可能造成偏差的经历和事件可以大致归入以下三类:长期趋势,短期事件和成熟过程。

长期趋势

在社区、地区或国家中,较为长久的态势通常用长期趋势来表达,长期趋势所带来的变迁可能会增进或者掩盖项目的外显效果。在社会生育率下降的阶段中,一个减低生育率的项目就可能表现为有效,其原因实际上极有可能是生育率下降趋势本身造成的偏差。同样,一个改善贫困家庭居住质量的项目也许比其实际效果表现得更加有效,因为国民实际收入增长的趋势使得每一个人都能在

其住房上投入更多的资源。另一方面，长期趋势也可能掩盖项目的真实效果。一个有效地促进粮食产量的项目，会由于项目实施期间不良气候的影响而得不到真实体现。类似地，一个为获释犯人提供就业机会的项目，很可能受到项目执行期间就业市场不景气的影响而变得效果不明显。

短期事件

与长期趋势一样，短期事件也可能造成一些意外变化，给对项目效果的估计带来偏差。比如，停电就可能中断通讯并影响食物的发放，于是，停电事件就可能影响某个营养项目的运作。同样，一场自然灾害可能使得增进社区团结的项目表现得有效，而实际上是突发危机使社区成员团结了起来。

成熟过程

前面提起过，影响评估必须常常应对这样的挑战，即自然的成熟和发展过程可能独立地对项目产生相当大的影响。如果把这些变化纳入项目效果的估算中，那么，估算就有偏差。例如，一项激发年轻人体育运动兴趣的社会项目的效果，可能会因为他们进入劳动年龄所导致总体兴趣下降而掩盖。成熟的自然过程也可能影响到年龄较大的成年人：在成年人中推进预防性医疗实践的项目可能看起来会没有效果，因为随着年龄增长，人们的健康状况总体上会不断下降。

非随机设计中的偏差

在所有并非完美实施随机实验的情况下，影响评估中的这些偏差对所有评估设计和分析而言，都是一个关键问题。事实上，在某些情况下，随机实验中也可能包含偏差。在随机实验中，产出测量没有损耗的完美随机过程应该避免选择性偏差。从随机分组到产出测量，精心地在项目组和控制组之间维持可比性，能够避免其他经历或者事件所造成的偏差。如果一个设计方案不能满足上述两个条件，那么，在对项目效果的评估中，就可能潜藏偏差。

在使用非随机化影响评估设计以及具有高度损耗或项目组和控制组受到不同外部事件及经验影响时，评估者必须关注估计中的偏差，并尽量减小偏差，这是研究设计的关键问题。因此，下面将围绕给项目效果评估带来偏差的影响因素以及降低偏差的方法，来讨论随机实验的备选方案。

准实验影响评估

第 8 章已经说过，准实验是指没有随机分配控制组的研究设计。在准实验设计中，接受项目干预的对象，与选定的、非随机指定的对象或没有接受实验干预的潜在对象构成的控制组进行比较。控制组在某些相关特征或经历上类似于项目组，或者可以通过统计手段对之进行调整以使两者类似，这样，就可以在相

当可靠性的基础上评估项目的效果。

准实验设计适用于可以利用随机方法却不能自始至终地保持其随机性的项目。举例言之,关于新奥尔良无家可归的药物或酒精滥用者项目,是一个为无家可归的酒精或药物滥用者服务的安居性成人再社会化项目(Devine, Wright, and Brody, 1995)。这一影响评估是作为随机实验设计的,但是真正随机分组的合适对象不到三分之一。项目人员改变了随机过程,他们只在观察到"好的"对象时,才运用随机结果,否则就置之一边。而且,人员减少以及进一步的资料搜集也破坏了最初的随机设计。由此产生的干预组和控制组不再符合随机实验的要求,因而评估者不得不将之视为准实验设计。

具体的准实验设计是否能对项目效果进行无偏估计,很大程度上取决于设计能在多大程度上减少控制组与项目组之间的关键差别。准实验研究的各组中常常存在着相关差别,因此也就存在对项目效果的估计产生偏差的可能。举例来说,假设有一个旨在促进农业生产的政府项目,通过广泛的教育运动指导农民如何通过合理施肥来增产。评估者通过寻问农业管理部门获知自愿参与项目的地区并展开评估;同时,还在没有自愿参与项目的地区挑选了一些与影响农业生产的特征(例如平均降水量、耕作面积、农作物和单位面积的平均投入)相匹配的地区作为控制组。难题在于,也许在两组之间还存在某些未知的因素,同样影响着农业产量。也许自愿参与项目的地区更加热衷于采用新技术或更加愿意承担风险。如果这些地区也采取了其他手段来增产,那么选择性偏差就会导致对项目效果的夸大。

因此,准实验设计要求对项目组和控制组采取严格的措施来保障两者的一致性。这里要介绍的准实验设计技术,包括通过配对来建构控制组,通过统计程序来均衡控制组与项目组,还包括回归—断点设计以及反身性控制的使用(在这种技术中,目标群体与其自身进行比较)。

通过配对来建立控制组

配对是准实验设计中建构控制组的一种方法。在配对设计中,常常是首先确定干预组,然后评估者通过一些特征的匹配来决定哪些对象可以和干预组相配对从而建构控制组。这种设计的逻辑是在两组都没有接受项目干预的情况下,任何可能导致产出差异的特征都是相匹配的。如果配对不能充分地平衡可能影响到项目产出的特征,那么对于项目产出的评估中就会存在偏差。

选择用以配对的变量

对于尝试使用配对设计的评估者来说,首要的困难就是了解哪些特征对于配对来说是重要的。评估者应该通过已有的知识和对相关理论的了解来决定。相关信息也可以通过项目关联领域中的研究文献来获取。例如,对于要减少未成年怀孕的项目来说,未成年少女怀孕的研究就有益于确定其性行为以及怀孕等行为的动因。建构一个匹配的控制组,就要尽可能选择与早孕重要特征相符

合的年轻人。

应该特别关注项目活动所针对的变量,即选择谁参与项目的变量。举例来说,就对失业年轻人进行职业培训的项目而言,年轻人对训练的态度以及对就业的评价就是匹配过程需要考虑的重要因素(Chen,1990)。即便选择分组的变量不能够完全匹配,评估者也应该确定并测量这些变量。倘若能够使之整合到资料分析中去,就更易于发现选择性偏差并对之进行可行的统计调整。

幸运的是,并不总是必须使用文献提供的每个影响产出的变量来进行配对。涉及的特征之间常常是相关的,因此有的因素在统计上就是冗余的,可以不加考虑。比如,教育干预的评估者可以基于智力水平对学生进行配对,也可以利用各年级学生的考试平均成绩进行匹配,因为智力测验的结果与考试成绩的相关性非常强。评估者应该对这些相关关系有所认识,尽量在各个有影响力但又不冗余的因素基础上进行匹配。如果不能区分影响产出的特征变量,两组的测量结果就会产生很大差异,那么,将会对项目效果产生有偏估计。

配对程序

配对的控制组,可以通过个体或整体的配对来建构。在个体配对时,需要在参与项目的学生群体中为每个参与对象抽出其“搭档”。例如,如果认为年龄、性别、兄弟姊妹数量以及父亲职业是相关的配对变量的话,那么在为干预组的每个孩子匹配最相似的孩子时,应该仔细考察参与项目的孩子名单。为了配对成功,可以适当调整近似的标准。例如,也可以在年龄相差 6 个月的范围对两组进行配对。专栏 9—C 提供了个体配对的说明。

专栏 9—C　利用个体配对控制组研究封闭式教育的效果

具有严重残疾的学生一直以来都是在封闭式特殊教育教室里学习的,但当前的政策讨论集中在是否在普通教育设施中安排更多的混合式教育。我们的兴趣之一在于,是否混合式学校能促进有严重残疾的学生的发展。混合式教育的合理性主要来自于这样的研究,研究表明在传统的特殊教育制度中,残疾学生和非残疾学生的社会关系非常弱。

为了评估混合式教育中残疾学生和普通学生社会关系的作用,研究小组用夏威夷瓦胡岛上的小学高年级学生设计了一个影响评估。在瓦胡岛,某些学校的残疾学生在普通教育的教室中接受特殊教育,然而在另外一些学校,残疾学生在具有特殊教育设施的封闭式教室中学习。

根据年龄、性别、残疾程度、适当的沟通能力以及适当的社会行为,来自于常规小学高年级课堂的 8 个残疾学生每一个都有一个处于特殊教育的学生来配对。统计分析显示两组学生之间没有显著差异。然后,根据学生之间的朋友网络以及他们与非残疾学生之间的互动特征,对两个组进行了比较。结果显示:在普通教育课堂中的学生与非残疾学生的互动范围更广、互动也更为频繁,其所得到和付出的社会支持水平也更高,而且有更广的朋友网络以及和非残疾学生更为持久的关系。

资料来源:Craig H. Kennedy, Smita Shikla, and Dale Fryxell, “Comparing the Effects of Educational Placement on the Social Relationships of Intermediate School Students With Severe Disabilities.” *Exceptional Children*, 1997, 64(1): 31-47.

对于整体配对而言,并不是一对一的个体配对,而是让干预组和控制组在每个配对变量的总体分布上一致。例如,我们发现干预组和控制组在孩子的年龄和性别上具有相同的比例,但这一结果可能来自于包括一个12岁女孩和8岁男孩的控制组以及包括一个9岁女孩和11岁男孩的干预组,而两组在总体分布上是平衡的。专栏9—D给出了一个整体配对的例子。

专栏9—D　利用整体配对的家庭发展项目评估

这个项目始于巴尔的摩,旨在通过整套措施为生活在公共住房的贫困家庭提供服务,希望能帮助他们摆脱长期的贫困。这些服务措施包括孩子和大人的特殊教育项目、就业培训项目、儿童项目、特殊健康护理以及便利的儿童看护等各种干预。尽可能在拉法耶特(LaFayette Courts)公共住房计划范围内提供这些服务。负责住房计划的管理者帮助不同家庭挑选适合于他们的服务。项目的显著特点在于,其重点是服务于家庭而不是个人。总共有125个家庭参加了这一项目。

为了建立控制组,从对应的一个名为墨菲(Murphy Homes)的公共住房计划中挑选了125个家庭。这样,通过与墨菲计划抽出来的家庭比较,来评估家庭发展项目的效果。一年以后,所有参与项目的家庭表现出更强的自尊以及控制自己命运的感觉,但是还未出现就业和收入方面的积极影响。

资料来源:Anne B. Shlay and C. Scott Holupka, *Steps Toward Independence: The Early Effects of the LaFayette Courts Family Development Center*. Baltimore, MD: Institute for Policy Studies, Johns Hopkins University, 1991.

个体配对通常比整体配对更好一些,尤其是当几个特征同时被用来作配对标准时。然而个体配对的缺点是费用大、费时而且很难对大量的配对变量进行操作。而且,个体配对有时会导致大量的数据损失。如果干预组之中的个体在控制组中找不到配对,就会被从数据库中剔除出去。在某种情况下,无配对个体的比例会很大,以至于有配对的个体不再具有实施干预的代表性。

补充一点,如果用来配对的变量可靠性低,就可能出现严重的统计假象。当用来配对的个体来自变量分布的不同状态时,这种情况尤为可能发生。假设有这么个例子,根据教师学术才能的排名来配对学生,这种配对就有一定的可靠性。如果一个补习计划的干预组来自于差学校且分数较低,而与其配对的学生来自于好学校且分数较高,很可能是从差学校来的高分学生与来自好学校的低分学生配对。如果根据这种情况下获得的结果进行组间比较,就可能出现与项目效果无关的伪差异(即对均值的回归;更多信息请参见 Campbell, 1996; Campbell and Boruch, 1975; Hsu, 1995)。

无论怎样小心配对,干预组和控制组之间总可能存在某些与结果相关的重要差异。如果相关变量已知且可测量,就可将其用作统计控制变量(将在下一部分讨论),尽管不能依此配对。实际上,在控制组配对设计中,一个有用的策略就是只使用部分重要变量来进行配对,对其他变量则进行统计控制。

实际而言,统计控制在近些年中已经在很大程度上替代并且完善了配对程

序。在许多操作中成为了配对的等价物。不过，至少在某些变量上，配对在影响评估中仍然较为常见。例如，牵涉到少数整体单位的比较，比如学校和社区。当目标对象是特定人群（例如具有某种医学的、私人的或者情景的特征的人）时，譬如在对医疗或相关健康干预项目的评估中，配对也是常用的方法（此时，配对通常被称作案例控制设计）。

通过统计程序来对等分组

在最常见的非对等比较设计中，项目组的产出被拿来与控制组的产出进行比较，而控制组是在相关性和便利性基础上产生的，且其相关建构方式与项目组有所区别。例如，为了给年长公民全社区干预项目建构控制组，评估者也许会从一个易于入手的社区获得控制组。任何基于对项目组和控制组产出简单比较的项目影响评估，都必须预设选择性偏差的存在。如果可以测量相关的原有差别，就可以运用统计手段来对组间差别进行控制，否则将会导致对项目效果的估计偏差。

为了解释统计控制背后的逻辑，我们将从下面这个简单的例子入手。专栏9—E是一个假设的准实验影响评估，这是一个针对35～40岁失业人员的项目。设计这种部分覆盖项目，是为了提高参与者的就业技能，使他们能够获得收入更高的工作。项目评估抽取1 000个参与者样本，并在他们完成职业培训之前和一年之后对他们进行访问。另外，在实施项目的大城市的总人口中抽取了同样年龄的另1 000个样本，在项目开始和结束一年之后对他们进行访问。两个样本都被问及当前的收入状况，并且从第二次访问中统计他们的小时工资水平。

在专栏9—E的第一部分中，比较了两组在没有任何统计控制情况下培训后的平均工资水平。参与项目的个体平均每小时挣7.75美元，而没有参加的个体是8.20美元。显然，项目参与者比非参与者挣得更少，如果这是一个随机实验的结果，那么这个差异应该是一种无偏估计。但实际上，除了是否参与项目之外，两组在许多与收入有关的变量上几乎都存在差异，显然，未经调整的比较包括了选择性偏差，并且可能会产生误导。

专栏9—E的第二部分引入了其中的一种差异，就是根据是否完成高中教育，对两种具有不同教育水平的人分别计算其工资水平。请注意，70%的项目参与者并未完成高中教育，而相应的非参与者中只有40%。当我们通过比较同等受教育水平者的工资水平，进而控制教育水平时，参与者和非参与者的小时工资非常接近，即未完成高中教育的分别是7.60美元和7.75美元，完成高中教育的分别是8.10美元和8.50美元。显然，控制同等受教育水平减少了参与者和非参与者的工资差异，因为受过高中教育的人通常获得更高的工资。

第三部分又引入了一个新的差异。由于所有参与者在其参加项目培训之初处于失业状态，因此，将参与者与项目开始时也一样失业的非参与者进行比较才是恰当的。在这一部分，非参与者根据项目开始之时的情况被分成了失业和未失业两部分。这样的比较表明，项目参与者在每一种教育水平上都比非参与者

要挣得多,即:未受过高中教育分别是7.60美元和7.50美元,受过高中教育者分别是8.10美元和8.00美元。总之,当我们引入了教育和失业两个因素之后,这个职业培训项目显示了某种积极的效果,项目参与者的小时工作增长了0.10美元。

专栏9—E 就业培训计划影响评估中的统计调整(假想的例子)

1.35~40岁的培训项目参与者和非参与者样本之间的产出比较

	参与者	非参与者
平均工资水平	$7.75	$8.20
n =	1 000	1 000

2.控制教育水平之后的比较

	参与者		非参与者	
	高中教育水平以下	高中教育水平	高中教育水平以下	高中教育水平
平均工资水平	$7.60	$8.10	$7.75	$8.50
n =	700	300	400	600

3.教育获得和培训项目开始时就业状况比较(或者说是非参与者的数据平衡)

	参与者		非参与者			
	高中教育水平以下,失业	高中教育水平,失业	高中教育水平以下,失业	高中教育水平以下,就业	高中教育水平,失业	高中教育水平,就业
平均工资水平	$7.60	$8.10	$7.50	$7.83	$8.00	$8.60
n =	700	300	100	300	100	500

在任何一个实例中,都可以加入一些控制变量,譬如先前的就业经历、婚姻状况、所要负担的人口数量以及种族等这些我们所知道的与工资水平有关的因素。即便如此,我们也不能保证对所有变量的统计控制能够从项目效果的估计值中完全剔除选择性偏差,因为控制组和项目组之间有影响力的差异可能仍然存在。例如,参与项目的失业者与未参与项目的失业者的差异恰恰是其更强的就业动机,这一差异是很难在我们的分析中得以测量的。

多元统计技术

专栏9—E中的调整,以一种简单的方式展现了统计控制的逻辑。在实际操作中,评估者通常采用多元统计方法来同时控制组间的各种差异。用这些方法产生统计模型,来表达控制变量与产出变量之间的关系。分析的目的在于,解释干预组和控制组之间的原始差异,通过减去完全属于原始差异的部分来调整其

最终差异。不论调整之后还剩多少差异，只要还存在着差异，就被当作干预的净效果。当然，如果模型中还有没有测量到的原始差异，如果这样的效果没有减去，那么剩下的差异也不是干预的真正净效果。这就是为什么在准实验影响评估中确定和测量所有与组间原始差异有关的变量是如此重要，因为这种差异与后来计算的产出变量净效果有关。

适用于非对等控制组设计的多元统计模型，一般包括一种或两种不同类型的控制变量。其中一种是与产出变量有关的成员特征。如果假定干预并不产生效果的话（或是根本就没有干预），那么正是这些研究之初测量的变量可以“预测”产出。例如，在专栏 9—E 中，教育水平就是这样的一个变量。如果其他条件都一样的话，在研究之初就具有更高教育水平的人最后也会得到更高的工资。类似的，如果没有任何干预作用的话，那些工作经验丰富的、社会地位优越的或在劳动力市场中比较抢手的人，也会获得更高的工资。

另一类型的控制变量，是与把成员选入干预组还是控制组有关的变量。这类控制变量与选择性偏差直接相关，而选择性偏差是准实验设计中的关键问题。这类控制变量包括如下几种：项目对象居住地离项目实施地的距离，参加项目的动力有多大，是否具有项目特征，诸如此类。这些控制变量的重要性在于，如果能充分说明把项目对象选入干预组的特征，就能对该特征进行统计控制，从而很好地抵消选择性偏差。

在下面的两部分中，我们将讨论相关的两种多元统计分析，首先讨论假设与结果直接相关的控制变量，然后是假设与选择相关的控制变量。后者无非是前者的一种特殊情况，但在程序上有很大的差别，因而有必要单独讨论。

建立产出因子模型

对非对等控制组设计的数据进行多元统计分析的目标，在于建立一个统计模型，根据每个人在研究之初的控制变量测量值来“预测”其产出变量值。如果干预组的平均产出优于根据其原始情况的预测，这种差异就可以被看作是干预的效果。换一种说法，我们的分析旨在确定是否在已经解释了控制变量的情况下，干预对于产出预测还是重要的。

基于这个目的的统计程序和统计模型有赖于各种测量值的特征、模型中所假设的变量间关系形式、认定的真实统计假设以及分析员的技术水平和偏好。通常情况下，所用技术与多元统计回归相似（Mohr, 1995; Reichardt and Bormann, 1994），有时也用结构方程模型（Loehlin, 1992; Wu and Campbell, 1996）。

专栏 9—F 是一项多元统计回归的应用实例，其数据来自于对参加以饮酒量为主题的酗酒者匿名（AA）会议的效果评估。专栏 9—F 中的表格是其回归分析的总结性统计图表，呈现了通过与饮酒量有关的项目变量预测项目活动后使用饮酒模式量表测量的得分。作为干预变量，把“参加 AA 会议”引入回归方程，通过使用控制变量来判断它对预测结果是否有显著贡献，是否超越了控制变量的

影响。

系数栏中的数值是非标准化的回归系数,表达的是每个变量增加一个单位对于饮酒模式量表测量得分增加的数量。参加了AA会议的系数是-2.82,意味着在其他变量相同的情况下,参加了AA会议的饮酒者比没有参加的饮酒模式量表得分要低2.82。因此,参加AA会议这一变量的回归系数就是要估计的项目效果,表明参加AA会议的酗酒者在会后要比没有参加的人更少饮酒。

在这个多元回归分析中,控制变量是挑选出来的,因为以往的研究表明:饮酒量与参加治疗项目或饮酒行为有关。因此,饮酒者的性别、信息搜寻、对酗酒严重性的认知、基本酒量以及婚姻状况都是在参加AA会议的效用之外对酗酒者的饮酒量发生独立影响的变量。专栏9—F显示出,在这些变量中,只有一个对于饮酒量的结果有显著性影响——基本酒量越大的人,更可能在产出测量时饮更多的酒。

专栏9—F展现的分析结果有多大意义,很大程度上取决于回归模型中使用的控制变量在多大程度上把握住了在饮酒问题上将人们区别为参加AA会议与不参加会议的那些因素。如果所有的差别都在回归模型中得到了体现,那么,是否参加AA会议对于产出数值的贡献值就是项目的效果。如果这些控制变量不足以解释选择性偏差,那么对项目效果的最终评估就会带有某种程度的未知偏差。

专栏9—F　使用回归模型估计参加酗酒者匿名会议的效果

参加酗酒者匿名会(AA)会议会影响酒徒们的饮酒量么?酒徒们的自愿参加正是AA哲学的一部分,因此自我选择就成了干预的一部分,若要通过比较AA会议的参加者与非参加者来评估其效果的话,必须先处理有关参与者特征的选择性偏差。帕罗·阿图退伍军人事务健康照料系统(Palo Alto Veterans Affairs Heath Care System)的研究者运用了包括简单多元回归模型在内的几种统计方法,通过统计控制建立对等组(控制组)。

首先考虑的是可以预测是否参加AA会议的变量。根据以前的研究,我们选择了三个变量:对酗酒严重性的自我认知、通过搜寻信息和建议以合作方式解决问题的倾向、性别。另外两个变量因为对饮酒量的作用也被纳入进来——基本酒量和婚姻状况。我们感兴趣的产出变量即饮酒量是通过饮酒模式量表测量的。

我们选取了218个有酗酒问题的个体作为样本来测量这些变量值,在我们建立的回归模型中,饮酒量是因变量,其他变量都是自变量。干预变量即参加AA会议(0=不参加,1=参加)也是预测变量之一,用来估算在其他变量得到统计控制的情况下,其对预测结果的影响。

如下表所示,有两个变量对产出有显著影响,其中包括干预变量即"参加AA会议"。"参加AA会议"的显著性负值系数表明,控制了模型中的其他变量之后,参加者比非参加者的饮酒量减少了。如果所有统计模型中的变量完全涵盖了选择性偏差,那么"参加AA会议"这个变量的非标准化回归系数就是对饮酒模式这个产出变量的项目效果估计值。

预测饮酒量的回归结果

预测变量	系 数	标准误
性别(0 = 男性,1 = 女性)	-1.16	1.09
信息搜寻	-.04	.12
对酗酒严重性的认知	-.44	.57
基本酒量	.20*	.09
是否再婚(0 = 否,1 = 是)	-1.69	1.25
参加 AA(0 = 否,1 = 是)	-2.82*	1.15
R^2 = .079		

* 表示统计显著性是 $p \leq .05$。

资料来源:Keith Humphreys, Ciaran S. Phibbs, and Rudolf H. Moos, "Addressing Self-Selection Effects in Evaluations of Mutual Help Groups and Professional Mental Health Services: An Introduction to Two-Stage Sample Selection Models." *Evaluation and Program Planning*, 1996, 19(4): 301-308.

建立选择因子模型

如专栏 9—F 所示的一步回归模型(one-stage)是根据控制变量与我们感兴趣的产出变量之间的关系来构造的。要使方程有效,必须整合每一个可能影响产出的变量和使干预组与控制组产生差异的变量值。而且,也可以将与干预组和控制组选择有关的变量纳入模型之中。例如,在专栏 9—F 中,研究人员选择了三个被认为与"参加 AA 会议"可能相关的控制变量——性别、信息搜寻和对酗酒严重性的认知。然而在回归模型中,是将其作为产出的预测变量而不是选择的决定因素来看待的。尽管与选择相关,但并不因此在方程中就能充分利用其所携带的具有选择影响的信息。

另一种越来越普遍的方法是利用两步(two-stage)回归分析。第一步是使用相关控制变量建立一个统计模型,以将个案分配到干预组或者控制组。第二步是使用分析的结果,将各个控制变量整合成一个判别变量或一个倾向评分(来判断被选择到哪一组)。在对我们关注的项目效果的评估分析中,倾向评分被当作一种整合的控制变量。在两步分析中,第一步的目的在于对个体分选到非随机的干预组或控制组进行统计描述。这就是所谓的**选择性建模**(Selective modeling)。

选择性建模有赖于评估者确定和测量相关选择变量,即与个体选择(如自愿)和被选择(如行政安排)过程有关的变量。因为所属组别是一个二分变量(如:1 = 干预组,0 = 控制组),适用于因变量为二分变量的回归模型常被用来进行选择性建模,例如罗吉斯蒂(logistic)回归。现在有几种不同的选择性模型以及对干预效果的两步估算。其中包括赫克曼的计量经济学方法(Heckman and Hotz, 1989; Heckman and Robb, 1985),罗森鲍姆赫鲁宾(Rosenbaum and Rubin,

1983,1984)的倾向评分(propensity sores)和工具性变量(Greene, 1993)。关于这个问题的一般性讨论,请参见汉弗里、弗布斯和摩斯(Humphreys, Phibbs and Moos, 1996)以及司徒恩博格和罗丽斯(Stolenberg and Relles, 1997)等人的论著。

专栏9—G是一个关于选择性模型的更详细的例子,这个模型解释了专栏9—F中酒徒们参加酗酒者匿名会(AA)的自我选择过程。研究者识别了三个以往研究文献中认为有可能与参加AA会议有关的控制变量。专栏9—G表明,这些变量用于第一步分析,预测参加AA会议的情况。我们运用这个分析的结果创造了一个新的变量λ,λ是第一步模型中各变量的最优拟合,我们用λ来判别酗酒者中谁将会参加AA会议。

源自第一步分析的变量λ,在第二步回归中被作为控制变量来预测结果,即饮酒模式量表测量的得分。为了达到这个目的,回归分析中又纳入了另外两个被认为有可能与饮酒量相关的控制变量。专栏9—G中展现的两步选择建模分析的结果,与专栏9—F中的一步分析建模的结果有明显差异。最明显的是,通过更好地控制AA参加者与非参加者之间的相关选择的差异,两步分析模型对项目效果的估计比一步模型在已有分析中估计的要大(用饮酒模式量表得到测量值是-6.31,而不是-2.82)。

专栏9—G　使用两步选择性模型估计参加酗酒者匿名会议的效果

通过单独估测选择效应对产出变量的影响,较之于专栏9—F中的一步多元回归分析,两步选择建模方法能对"参加AA会议"的项目效果作出更好的估计。研究者认为有以下三个可以预测是否参加AA会议的变量:对酗酒严重性的认知(认识到酗酒问题的人更可能参加)、通过搜寻信息和建议以合作解决问题的倾向、性别(女性比男性更可能寻求帮助)。这些变量在第一步分析中被用来预测是否参加AA会议,而不是像在一步模型中那样,直接用来预测产出。正如下表所示,其中有两个变量与"参与会议"有显著的独立性关系。

第一步:预测参加AA会议的概率回归分析

预测变量	系　数	标准误
性别(0=男,1=女)	.29	.19
信息搜寻	.06*	.02
对酗酒严重性的认知	.38*	.09
R^2 =.129		

*p≤.05

然后,用选择性模型来建构一个新的变量λ,λ表达了每个个体参与到干预组而不是控制组的可能性。在第二步回归中λ被用作控制变量来预测产出,即根据饮酒模式量表测出的饮酒量。这一步也纳入了另外两个与产出相关的控制变量,即基本酒量和婚姻状况。最后,纳入干预变量"参加AA会议"(0=否,1=是),我们可以对包括选择变量在内的其他预测变量进行统计控制,进而评估干预变量与产出变量的关系。

"参加AA会议"的显著性系数表明,控制了基本酒量和自我选择之后,参加者比非参加者的饮酒量减少了。事实上,在使用总分为30分的饮酒模式量表测量后,"参加AA会议"的净效果至少使之降低了6分。请注意,使用两步模型得到的对参加AA会议效果的估计值比使用一步模型增加了将近一倍。

第二步:预测最终饮酒量的最小二乘法回归

预测变量	系 数	标准误
基本酒量	.20*	.08
已婚(0=否,1=是)	-1.68	1.23
λ	2.10	1.98
参加AA会议	-6.31*	3.04
R^2 = .084		

* $p \leq .05$

资料来源:Keith Humphreys, Ciaran S. Phibbs, and Rudolf H. Moos, "Addressing Self-Selection Effects in Evaluations of Mutual Help Groups and Professional Mental Health Services: An Introduction to Two-Stage Sample Selection Models." *Evaluation and Program Planning*, 1996, 19(4): 301-308.

由第一步分析所得到的整合的选择变量也可以用作配对变量,在干预组和控制组中再建立子群体。这种方法被称作倾向评分分析(propensity score analysis)(Rosenbaum and Rubin, 1983,1984)。这一分析通常将倾向评分选择变量值的分布进行五等分(分成规模相等的五个组),每一个五等分群体都包括了倾向评分相近(在每一个五等分范围之内)的干预组和控制组成员。因此,可以对每个五等分群体独立评估项目的效果,并最后形成一个整体性的估计。

相对于专栏9—F和9—G的两步多元回归方法,倾向评分配对技术的优点是:统计假设更少。但这种方法调整选择性偏差的能力仍然严重依赖于对影响分组变量的确认,以及将其纳入到产生倾向评分统计模型的能力。

回归—间断点设计

前面关于选择性模型的讨论应该已经阐明,影响到是否纳入准实验干预组和控制组的基本变量数据如果充分有效,就能够建立非常有效的统计控制。假设现在无须知道什么变量与选择有关,而是直接给评估者列出选择变量,然后挨个计算项目对象在这些变量上的得分,再据此将其分成干预组和控制组。在这种情况下,由于没有任何关于选择的不确定因素,而且评估者手头掌握着决定模型的已测数据,那么肯定能建立一个选择性模型。

结构化控制组准实验的一种特殊情况即我们所谓的"回归—间断点设计",就是基于这样的情况。尽管这种设计的应用受到某种程度的限制,但一旦被运用,却可以获得比其他准实验影响评估设计更小偏差的项目效果估计。回归—间断点设计适用于这样的情况,即评估者无法将对象随机分配到干预组但却能

与项目人员合作将对象基于其需求、价值或其他情况进行系统分配,可以把最有需求、最有价值等特点的个体指派到干预组,而其他人则作为控制组。

回归—间断点的一个更形象的说法就是临界点设计,因为这种方法是对一些诸如需求、价值等具有连续性特征的变量设定一个临界点,得分高于临界点的人则被归为一组(例如控制组),得分低于临界点的人则被归为另一组(例如干预组)。由于是根据一个可测量变量的得分来选择参与者的,从而选择过程是清晰明确的,这个可测量的变量也使得对选择性偏差的统计控制变得相对简单。在已知的选择过程有一个恰当模型的情况下,才在产生对项目效果的无偏估计方面使得回归—间断点设计近似于随机实验。

例如,为了评估加利福尼亚刑满释放人员获取失业保险资格认证的项目(有些类似于专栏 8—F 所描述的巴尔的摩 LIFE 实验),伯克和骆马(Berk and Rauma, 1983)利用了项目中的资格认证(主要依据罪犯在监狱中的工作天数)。这些释放人员必须在狱中工作满 652 天才有资格获得失业保险,而且保险金额是按照工作天数进行比例划分的。这个公式是明确的,而且是定量的,保险金获取资格的临界点一律是工作 652 天,高于这个点就给失业保险,低于这个点则没有。

将获得保险金与没有获得保险金的再犯率比较,若以狱中工作时间为控制变量(即建立选择性模型),可以估计保险金对于再犯的作用。回归分析表明,获得保险金的获释人员再犯率要低 13%。只要选择过程是已知的,且在统计分析中是以变量(狱中工作时间)来表述的,那么这一估计就是无偏的。

尽管有某些优势,但临界点设计却不经常使用,部分原因在于并不是所有项目都按照评估者希望有"资格认证"那样明确的、精确的规则或者愿意采用这样的规则。然而,另一个原因可能是评估者本身对这种方法理解不够充分,因此,即使在适合运用的情况下,也常常没有采用。无疑,进行项目影响评估的评估者应该能从学习这些设计中获益匪浅(有关此类设计的资料有 Trochim, 1984; Braden and Bryant, 1990;Shadish, Cook and Campbell, 2002; and Mobr, 1995)。

反身控制

在使用反身控制(reflexive controls)的研究中,对于项目效果的估计完全来自干预对象在两个或者更多时点上的信息,这些时点中至少有一个处于项目开展之前。在使用反身控制时,除了干预变量引起的改变,必须确定干预对象的产出变量在观察过程中没有改变。如果是这样,项目干预前后产出状况的任何差别都可以被认为是项目的效果。假设某大公司原来通过邮寄支票来发放退休金,现在则自动存入退休人员的银行账户。如果邮箱遭窃的发生率、邮政服务的水平等条件没有发生变化,那么比较一下这个程序执行前后项目对象对发放延迟和错漏的抱怨,就可以分析出这个程序变化作用的显著程度。这是一个简单的事前—事后分析例子,接下来我们要描述其基本程序。随后,还会讲到反身控制设计的典型类型,即时间序列设计。

简单事前—事后研究

在一个简单的事前—事后设计(或者事前与事后研究)中,进行产出测量的

是同一组干预对象，在参与项目前做一次，在项目进行到足够产生一定效果的一段时间之后再做一次。然后，对这两次测量结果作比较，就能提供对项目效果的估计。这种设计的主要缺点在于：如果在前后测之间有其他因素对测量结果产生影响，那么，对于项目效果的评估就是有偏的。例如，可以通过比较参与医疗保险制度前和几年之后同一批人的健康状况来评估医疗保险制度的效果。但是，这种比较很可能会引起误导——人们自身的衰老会使得健康状况整体性降低，这将会使得对项目效果的估计带有向下的偏差。在个体参与医疗保险期间，其他生活变化例如退休和收入减少，也可能影响健康状况。

有时候，时间性的变化是不易被察觉的。例如，在对抑郁症医治方法的效果研究中，运用反身控制是否合适就值得考虑。当人们觉得情绪低落时，更容易寻求治疗；后来，症状很可能会自然地减轻，会不再觉得那么抑郁。因此，对抑郁状况的前后测，即便在治疗没有发生积极效果的情况下，也总会表现出状态的某种改善。

总的说来，简单事前—事后自反设计将产生对项目效果的有偏估计，因而对影响评估来说没有太大价值。尤其是当两次测量的时间间隔太长时——比如说一年或一年以上——因为随着时间流逝，其他过程更可能使项目的作用模糊化。因此，简单事前—事后设计主要适合旨在对基本条件不会变化的项目进行短期影响评估。在第 7 章也说过，简单事前—事后设计也适宜用于对日常产出的监测，其目的主要是为项目管理者提供信息反馈，而不是对项目效果进行可靠估测。

如果可以在项目实施前后的时间跨度内对产出进行多次测量，简单事前—事后设计的作用就能够得到加强。在这个时间序列中进行的反复测量，使人们有可能描述出当时正在发展的、使事前—事后效果评估产生偏差的趋势与因素，并用于修正对项目效果的有偏估计，这就是时间序列设计的基本理念。下面我们将简要讨论时间序列设计。

时间序列设计

时间序列设计这种反身控制设计，包括在整个干预期间多次进行的观察。举例来说，对退休金发放的反馈不仅仅是前测、后测各一次，而是支付程序变更两年前到变更一年之后每个月测量一次。在这种情况下，对项目效果的确定程度将会提升，因为有更多的信息用来估计在支付方式没有发生变化的情况下事情将会怎样。第二步是将产出数据依据对象的不同特征进行分解。例如，通过考察犯罪率高低不同地区以及城乡地区支票收取的不同状况，来进一步分析对养老金发放抱怨的时间序列数据，就会得到养老金发放程序变更影响的更多洞见。

在时间序列设计的每次测量中，被测量的对象可以是同一些个体，也可以不一样。时间序列方法经常从与研究结果（例如生育、死亡和犯罪）相关的、可以得出阶段性信息的既有数据中抽取数据。可用的数据通常包括总体性数据，例如为某种政治目的而计算出的均值或者比例。譬如，美国劳动部（Department of Labor）就保存着很好的时间序列数据，即 1948 年以来美国全国以及各个主要地区每个月的失业率。

当通过采用干预前的观察建立了一个较长的时间序列时，就可以针对目标群体建立长期趋势模型，考察干预持续期间甚至更长时间内的趋势，且预测干预后某

段时期内是不是有显著差别。运用这种“差分自回归滑动平均模型”(auto regressive integrated moving average,ARIMA,又被称作“求合自回归滑动平均模型”,Hamilton,1994;McClearly and Hay,1980)的总体时间趋势建模程序,可以通过考察长期趋势和周期性变量用以识别最优统计模型,也可以处理某一时点的取值关联于此前取值的情况(技术上被称作自相关)。这是一个技术含量很高的工作,并且需要对统计知识相当熟悉。

专栏9—H展现了一个使用时间序列数据的例子,人们用时间序列数据来评估提高法定饮酒年龄对减少酒后驾车引起交通事故的效果。正因为采用了较长时间序列的测量,才可以对多产出变量(超过200个)进行评估。这项分析通过收集不同合法饮酒年龄人群在降低酒后驾车交通事故政策出台前8~10年的事故率数据,与根据先前趋势预测的合法饮酒年龄提高之后的事故率增加进行比较,提供了对项目效果的测量。

专栏9—H 依据时间序列数据评估提高饮酒年龄的效果

20世纪80年代,美国许多州都将最低饮酒年龄从18岁提高到21岁,特别是联邦1984年的统一饮酒年龄法案通过之后,这项法案削减了对最低饮酒年龄在21岁以下各州的公路建设资助。改变的原因是基于这样的观念:较低的饮酒年龄导致了更多的十几岁青少年酒后驾车,并造成悲剧的增长。然而,对于提高饮酒年龄的效果评估却较为复杂,因为新型汽车安全技术的引入以及公众对酒后驾车危险性的认知,都会使得事故发生率呈现下降趋势。

威斯康星州1984年将最低饮酒年龄提高到19岁,随后1986年又提高到21岁。为了评估这一变化的影响,菲格利(Figlio)通过年龄分层,考察了18年中每个月酒后驾车造成交通事故的时间序列数据。数据来自威斯康星交通部门(Wisconsin Department of Transportation),时间跨度为1976—1993年。时间序列统计模型分别针对18岁(1984年以前可以合法饮酒)、19岁和20岁(1986年以前可以合法饮酒)以及21岁以上(始终可以合法饮酒)样本的数据。产出变量是各个年龄组每1 000个持照驾驶者中因饮酒引起的撞车率。

结果表明:对于18岁组而言,将饮酒年龄提高到19岁,使得因饮酒引起的撞车在此前的每千人每月平均2.2起的基础上减少了26%。对于19岁和20岁组来说,将饮酒年龄提高到21岁,使得因饮酒引起的撞车在此前的每千人每月平均1.8起的基础上减少了9%。作为对比,法案变化对于21岁以上组而言,效果只有2.5%,并且在统计上并不显著。

评估者的结论是:威斯康星州提高饮酒年龄法案的实施具有迅速且确切的效果,较之于法案实施前的平均状况而言,十几岁的年轻人酒后驾车事故有了实质性减少。

资料来源:David N. Figlio,“The Effect of Drinking Age Laws and Alcohol-Related Crashes:Time-Series Evidence From Wisconsin.” *Journal of Policy Analysis and Management*, 1995,14(4):555-566.

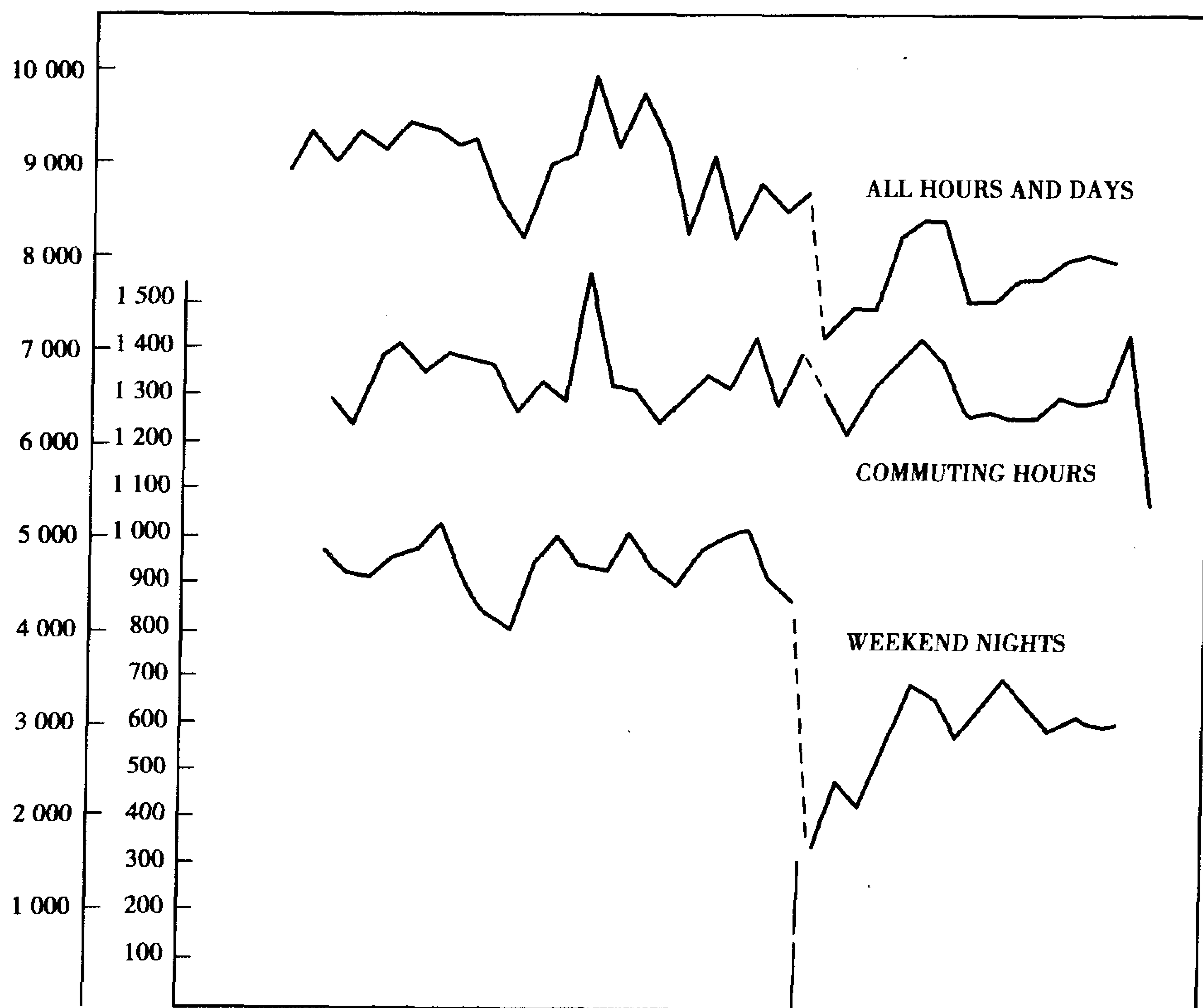
前文已经提到过,社会项目时间序列数据的分析单位通常是高度总体性的。专栏9—H中的例子是威斯康星州的,事故测量整合了全州的相关数据,且以每1 000个有照驾驶者中事故率的形式表达出来。与我们讨论过的其他类似模型一样,适用于此类数据的统计模型容易带有偏差。举例来说,如果在威斯康星州,存在有模型没有表达出的在涉酒事故上的其他显著性影响,那么分析的结果就不真实。

描述干预前后时间序列数据的简单图示方法，可以对我们了解项目的效果提供粗略但有益的线索。如果混淆在干预效果上的干扰性影响已知，且有很大把握确认其作用很小，那么，通过对时间序列图的简单考察就可以确定项目的显著效果。专栏9—I展示了在项目评估中时间序列原始数据的一个经典应用——英国的酒精测定条例(Ross, Cambell and Glass, 1970)。该图展示了酒后驾车罚金大幅调整的法案实施前后英国车祸的发生率。随附的图表说明，该法案具有可识别的干预效果：法案生效后，事故明显减少，尤其是在消费酒类较多的周末。尽管从图表上看效果是明显的，但是最好再用统计分析方法来确证一下，而专栏9—I所示的事故减少，实际上，在统计上也具有显著性。

专栏9—I　强制酒精测试对交通事故影响的效果分析

1967年，英国政府实施了一项新政策，允许警察在交通事故现场进行酒精测试。该测试旨在测量嫌疑人血液中的酒精含量。同时，对于酒后驾车者设定了更高的罚金。对这项从1967年10月开始实施的新法案，英国政府进行了大量的宣传。

下图展示了新法案实施前后一周之内不同时段的车祸发生率。对于下图的直观考察可以发现，在法案实施后，事故明显减少，一周内绝大多数时间都是如此，尤其是在周末时段出现了急剧的下降。同时，统计检验确证了较之于偶然因素所带来的影响，法案实施导致的降低更为明显，具有统计显著性。



资料来源:Summary of H. L. Ross, D. T. Campbell, and G. V. Glass, "Determining the Social Effects of a Legal Reform: The British Breathalyzer Crackdown of 1967." *American Behavioral Scientist*, 1970, 13 (March/April): 494-509.

时间序列方法并不只限于对单一案例的研究。当时间序列数据包含不同时间地点的不同干预因素时,也可据以进行更加复杂的分析。例如,帕克和雷布汉(Parker & Rebhun,1995)运用1976—1983年间美国50个州以及哥伦比亚特区的时间序列数据,考察了设定买酒最低年龄的国家法令变化与杀人率变化之间的关系。他们使用了集合的截面时间序列分析(pooled cross-section analysis),用虚拟编码(0和1)来代表是否达到了法定购酒年龄。该模型中的其他变量包括酒精消费量(人均消费啤酒的桶数)、婴儿死亡率(作为贫困指标)、不平等指数、种族构成、地区以及各州总人口。这个模型用于研究不同年龄群体的死亡率。研究发现,提高允许买酒的年龄与21~24岁年龄组的杀人事件受害者数量减少之间有显著关联。

尽管我们讨论过的时间序列分析都使用总体数据,但同样的分析逻辑也适用于非总体数据。例如,对一些小群体进行干预效果分析,在参加项目的前后甚至在项目进行过程中,对这些人的行为进行多次测量。例如临床医学家通过时间序列设计来评估治疗措施对患者的效果。因此,用成绩测试来测量儿童的表现,也许要在新的教学方法使用前后都进行阶段性测量;对成年人酗酒行为的测量也要在针对酗酒治疗开展前后都进行。当运用于单一案例时,适用的统计方法会有所不同(Kazdin, 1982),长期趋势和周期性问题对于单独案例来讲并不是那么突出,但是时间序列分析的逻辑仍然是一样的。

在影响评估中运用准实验方法的注意事项

执行良好的实地实验在项目效果无偏估计上的科学可靠性,使得这种设计成为最方便易行、最经济实惠的选择。不幸的是,社会项目的大环境使得随机实验难以实施。准实验设计的价值在于,如果谨慎运用的话,即使与项目环境只是相对适应且与社会研究的要求并不具有内在一致性,也可以对项目的净效果作出较可信的估计。总之,对于项目评估而言,准实验研究设计的好处在于,当随机实地实验不可行时,准实验设计方便并可行。

关键在于,好的准实验设计如何才能对项目净效果作出有效的估计。换句话说,就是基于现实原因运用准实验设计来取代随机实验的时候,评估者在评估过程中犯严重错误的风险有多大?我们希望引用一项比较研究来回答这个问题,即对不同项目环境中随机实验与多种准实验的结果进行比较。然而这样的研究很少,所以我们很难拿出确凿的证据来证明我们的判断。然而,仅有的一些例子已经说明:在良好的项目环境和谨慎的操作下,准实验研究设计既能够得出与随机设计相当的项目净效果,也可以谬之千里。

弗雷克和梅拉德(Fraker and Maynard, 1984)比较了配对建构控制组和随机实验两种方法获得的对一项就业计划效果的估计。他们利用不同的方法进行个体配对,但没有一个结果与随机实验的结果非常接近。但是,赫克曼和霍兹(Heckman and Hotz, 1989)使用同样的资料,采用更合适的控制变量和更复杂的统计模型,却得到了类似于随机实验的结果。类似的是拉隆德(Lalonde, 1989)对另一个就业培训项目的比较。在研究中,拉隆德拿非随机设计的统计控制结果与随机实验的评估结果进行比较,也发现了致命的矛盾,其中之一就是女性(在准实验中其作用被低估)与男性(被高估)的偏差。

最近,来自于不同项目领域中的证据已经有所进展。艾肯(Aiken et al., 1998)等人在一项大学新生拼写纠正项目中,比较了不同影响评估设计的结果,他们发现:在利用统计控制的非对等(非随机)比较设计和回归—间断点设计以及随机实验中,项目效果估计"高度相似"。然而,在简单的事前—事后比较中,却高估了项目净效果。

在同一项目领域中,一系列独立影响评估研究的广泛比较也说明了结构化控制的有效性。李普希和威尔森(Lipsey and Wilson, 1993)在心理学、教育学及行为学研究的74个事后分析中,比较了随机实验和非随机实验的平均效果。在许多包含事后分析的研究中,非随机实验的效果估计值与随机实验所得值非常近似。然而,也有同样多的案例表明两者有重大差异。海茵兹曼和谢德西(Heinsman and Shadish, 1996)曾对4个项目领域中的98项研究的效果估计值作了更精确的比较,得到了同样的结果:非随机设计与随机设计的比较有多种结果——有时候两者差不多,有时候大一些或小一些。

这些比较的结果表明:在一个既定的应用项目中,运用结构化控制的项目评估所产生的项目效果估计值与随机实验中的结果是类似的,但两者也经常会有不小差异。而且这些经验研究也并不能说明对于准实验设计而言,什么样的项目环境或什么样的设计会得到更好的或是更差的结果。因此,评估者在项目评估中运用非随机设计时必须根据具体的假设、已经选定了的设计要求、项目特征以及对象总体属性逐个进行分析,这样才有最大可能得出有效的项目效果评估值。

在本章已经考察过非随机效果评估设计的所有局限性之后,使用这些方法是不是就变得可接受了呢?很显然,在可以运用随机设计的时候,就不应该运用非随机设计。然而,当随机设计不能应用而又有必要进行效果评估时,如果评估者能够认识到非随机设计的局限性,并且努力克服这些局限性,那么评估者仍然可以应用非随机设计来进行研究。

负责任的承担一项只能用非随机设计方式完成的评估任务的研究者,有责任向项目方事先说明这种方法对项目效果的评估结果并不是确切无误的。不过,一旦决定进行非随机设计,研究者就应该回顾相关研究文献,以搜集可用于统计控制和选择性建模的变量。同时,在报告非随机影响评估的结果时,评估者也有义务表明对于结果的估计可能是有偏的,并尽可能指明偏差有多大。

小　结

- 效果评估的目的在于判别产出的哪些变化可以归因于待评估的项目干预。在此目的下,最有力的研究设计就是随机实验,当随机将目标对象指定到干预组和控制组不可行时,也可以采取几种准实验效果评估方法。
- 在运用任何一种影响评估设计时,评估者的主要关注点在于减少项目效果评估中的偏差。在各种偏差来源中,对准实验研究来说最成问题的是选择性偏差、长期趋势、短期事件和自然成熟。
- 在准实验研究中,干预组和控制组是由随机分派之外的方法建构的。准实验设计背后的逻辑与随机实验基本相同,除了在研究之初干预组和控制组不是对等的以外。在没有干预的情况下,两组之间的差别如果导致不同的产出,在对于项目效果的评估中就会包含偏差。因此,在准实验设计中,必须采用恰当的方法调整在项目净效果评估中产生的偏差。
- 有一种准实验设计需要用到配对控制。在这样的设计中,通过将项目参与者与未参与者(可以是个人或群体)配对来建构项目组。为了避免源自设计的项目产出评估偏差,两组对象据以配对的变量必须尽可能包含所有与产出强相关的因素。
- 干预组与控制组也可以通过统计程序(统计控制)来达到对等。同样,必须识别两组之间存在差别的、与产出相关的任何变量,并纳入统计调整范围。在这种设计中,人们通常使用多变量统计方法来同时控制多种组间差别。多变量统计分析会用到假定与产出相关的控制变量(建立产出因子模型)或用到在进入控制组还是干预组选择中发挥作用的控制变量(建立选择因子模型)。
- 当有可能以目标对象在需求、价值之类的某种量度上的测量为基础来决定被分入干预组或者控制组时,较之于其他准实验设计,运用回归—间断点设计来评估项目效果,会更少受到偏差的影响。
- 反身控制设计,既包括简单事前和事后比较,也就是在项目前后各进行一次测量;也包括在项目前后进行多次测量的时间序列设计。时间序列设计通常比简单事前—事后设计能更好地估计项目效果。
- 当随机设计不可行时,评估者可以使用准实验设计来进行效果评估。但是,必须努力减少可能产生的偏差,并且充分地意识到准实验设计的局限所在。

基本概念

损耗(Attrition):输出数据的损失,由对象属于控制组还是干预组来衡量,通常是因为对象流失或不能提供数据产生。

配对(Matching):寻找与干预组在相关方面完全一致的(单个的或群体的)对象用以建构控制组。

非对等比较设计(Nonequivalent comparison design):在准实验设计中,通过非随机分组方法建立控制组和干预组。

事前—事后设计(Pre-post design):反身控制设计中,在干预的前后均只进行一次测量。

反身控制(Reflexive controls):在项目干预之前对参与对象的产出变量值进行测量,并将其作为对照标准(控制)观测值。也可参见事前—事后设计;时间序列设计。

回归—间断点设计(Regression-discontinuity designs):准实验设计的一种,根据某变量的观测值范围是高于还是低于临界点来选择项目组或者控制组。也叫做临界点设计。

选择性偏差(Selection bias):由项目组和控制组之间不可控差异导致的对项目效果的系统性高估或者低估,会使各组在没有受到实验因素干预的情况下,也存在产出差异。

选择性建模(Selection modeling):在一个非对等比较设计中,建立多元统计模型以"预测"选入干预组或控制组的概率。分析的结果用来构建一个针对选择性偏差的控制变量,此变量可以为研究项目干预效果的第二阶段统计模型所用。

统计控制(Statistical control):运用统计技术,针对反身控制组和项目组之间与项目产出相关的差异,调整对项目效果的估计偏差。通过控制变量的引入,两组对象之间的差异可以在统计分析中被呈现出来。

时间序列设计(Time-series design):一种建立在干预前后对产出变量进行多次重复测量基础上的反身控制设计。

10 探明、解释和分析项目效果

上面三章主要讨论了项目产出的测量，关注项目效果评估的研究设计。尽管好的测量和设计方案不可缺少，但是一个项目的实际效果并不必然按照项目预期的形式发生。这就要求评估者关注项目的实际运作，关注项目效果的大小乃至存在与否。而且，即使可以得到一个很确定的评估结论，评估者也不能忽略项目效果的实际意义。

这一章介绍一些有关项目效果研究和解释的方法，分享我们在这方面的一些思考。本章还将讨论表示和呈现项目效果大小的方法，展现可能会影响项目效果的各种因素。对这些问题的了解，有助于评估者设计出更好的效果评估方案，也有益于促进社会项目效率等方面知识与经验的增长。

影响评估的最终产品是有关项目效果的一系列估计和评判。评估者通过比较项目参与者的产出和非参与者的相应产出，来分析和估计项目的效果。正如第8,9章所讨论的，在评估项目参与者产出方面，研究设计的可信度有很大的差异。然而，所有的影响评估，包括通过随机实验方法进行的评估，都有必要接受进一步的检验，以确定其实际功效。如何做好这一评估工作就是本章内容的核心。我们会考虑评估者如何才能弄清楚具体项目效果的大小，讨论如何评估这些效果的实际功效，分析如何在资料和数据中探寻项目效果。然后，针对不同项目对象，我们还会讨论项目效果在次级目标群体中的差异，这是一个更为复杂的问题。本章最后，我们还会简要讨论项目效果的"事后分析"（经常在影响评估中可以看到）如何有助于改善具体的项目设计和分析，有益于丰富和发展评估领域的相关知识。

项目效果的大小

影响评估可以探寻和描述项目效果，这种能力的发挥很大程度上依赖于项目实际产生效果的大小。与较大的项目效果相比，较小的项目效果当然更难于探明，同时，它们的实际意义也难以描述。本节主要讨论发现和描述项目效果，要做到这一点，就要认真思考——具体干预项目的效果大小究竟意味着什么？

在影响评估中，项目效果主要展现的是干预组和控制组在测量结果上的差异。干预组是接受了项目干预的目标人群，而控制组则是没有受到项目影响的另一个群体。因此，描述项目效果最直接的方式，简单说来，就是比较这两组对象在产出测量上的数量化差异。例如，一个公共健康运动项目可能意图劝服有高血压的病人去进行血压检测。如果针对目标对象的调查显示出，干预组过去六个月的检测比例为0.17，而控制组的相应比例为0.12，那么，项目效果就是0.05的比例上升。类似，如果对接受过项目干预的人进行高血压知识的多指标测量，其平均得分为34.5，而未接受项目干预者的平均得分为27.2，那么体现在这个测量工具上的项目效果，就是高血压知识量上7.3的增长。

以这种方式来描述某个项目的效果大小，对于上述目的是有效的。但是，用来估计项目产出的专门测量工具都非常特殊，不一定对所有对象都完全适用。譬如在知识竞赛中，用被试者熟悉的专门量表测到知识量有7.3分的差异是有意义的，如果对不熟悉量表的人进行同样的测量，其结果就没有什么意义。为了对项目效果大小进行一般性的描述，或为了运用统计方法展现项目效果，最好的方法是不使用特别专门的测量工具或测量程序。

显示某个项目效果整体大小的通用方法是用百分比的上升或下降进行描述。譬如，对于促使更多的老年人参加血压检测的干预项目来说，从0.12到0.17的比例上升代表了41.7%的数量增长(0.05/0.12)。必须注意的是，某一测量维度上百分比的上升或下降，仅对具有绝对零点的测量才有意义。所谓绝对

零点,就是指可以测到的零值。如果所研究的群体在6个月内没有任何人参加血压检测,其比例就是0,这是一个典型的绝对零点。因此,41.7%的增长对描述项目对象的变化是有意义的。

相反,用量表测量针对高血压的知识就没有绝对零点。如果因为知识晦涩难懂,那么测得零分的人可能仍具有可观的知识量。例如,有些老年人知道很多关于高血压的知识,只是不能以“心脏收缩”、“钙质抑制剂”这类概念进行完整的定义。如果测量工具包含许多这类测量指标,就会低估目标对象的知识储备,混淆真实的项目效果。另外,也可以用这种方式建构测量尺度,即最低分数并不必然为0,也可以为10。依照这种测量规范,描述干预组增加了7.3分就是无意义的,因为27%的知识增量仅仅是因为34.5在数字上比控制组的得分27.2高出了27%而已。如果测量规范和标准不同,同等的实际知识量差异,即使在控制组75分基础上增长10分,也只是增加了13%。

因为许多产出测量尺度都没有绝对零点,所以,评估者经常使用**效果大小统计量**(Effect size statistic)来描述某个项目效果的大小,而不是用原始结果得分或是简单的百分比变化。效果大小统计量要表达的是基于标准化形式的项目效果大小,这种方法可以使具有不同测量单位的结果之间具有可比性。

最常见的、以数字形式表示项目效果的统计量是**标准化均值差**(Standardized mean difference)。标准化均值差是干预组与控制组以标准差为单位的结果平均值差异。标准差是描述个体或者其他单元之间变异性的指标,在给定测量条件下,提供关于测量结果数值范围或者分布的信息。因此,标准化均值差用来描述项目效果的大小(以标准差为单位),显示与研究得到的最低值和最高值之间的测量值范围的相关程度。假设把对阅读速度的测试检验当作对学前教育项目的影响评估,测量结果显示:干预组的平均得分比控制组高出了半个标准差。这种情况下,标准化均值差影响大小是0.50,如果一个测量词汇量的标准化均值差是0.35,我们就可以直接拿这两个值进行比较。结果显示:学前教育项目在提高阅读速度方面更为有效,而提高词汇量方面的作用则相对较弱。

不过,某些测量结果是二分的,而非等级性或连续性的数值。也就是说,某个参与者或经历了一些变化,或没有。二分的例子很多,如实施违法行为、怀孕、从高中毕业等。对于二分变量,经常用**发生比**(Odd ratio)来表示项目的效果大小。发生比显示的是干预组与控制组相比较,在某个结果性事件中有多大的发生比例之比。1.0以上的发生比意味着干预组有更大改变的可能。譬如2.0的发生比就可能意味着干预组比控制组成员有多出一倍的可能性产生预期的项目产出。而发生比小于1.0,则意味着干预组有更少可能获得项目的预期产出。

这两种表示影响大小的统计量,在专栏10—A中都有较详细的说明。

专栏10—A 通用的效果大小统计量

标准化均值差

标准化均值差尤为适用于测量在持续干预下的项目效果大小,也就是说,适合测量有一定持续性的项目效应。可持续测量的变量包括年龄、收入、住院天数、血压记录、检测得分和其他这类标准

化的测量维度。这类测量结果以干预组和控制组之间的平均值形式呈现出来，平均值之间的差异正好显示出项目干预效果的大小。表示效果大小的标准化均值差如下：

$$\frac{\bar{x}_i - \bar{x}_c}{sd_p}$$

其中：$\bar{x}_i$ = 干预组的平均分数；$\bar{x}_c$ = 控制组的平均分数；

sd_p = 干预组的标准差 sd_i 与控制组标准差 sd_c 的总体均值。

$$sd_p = \sqrt{((n_i - 1)sd_i^2 + (n_c - 1)sd_c^2)/(n_i + n_c - 2)}$$

其中，n_i，n_c分别是干预组和控制组的抽样规模。

因此，标准化均值差代表了以标准差为单位的干预影响。按照常规做法，当项目产出更有利于干预组时，这一效果大小统计量就是正值；相反，如果对控制组更为有利，那么就代表了一种负向的项目效果。例如，如果对干预组来说，一个环境态度的平均得分是 22.7（$n = 25$，$sd = 4.8$），而控制组的相应得分是 19.2（$n = 20$，$sd = 4.5$），那么，干预组的较高得分就代表了一种积极正面的项目效果，其影响大小将是：

$$\frac{22.7 - 19.2}{\sqrt{((24)(4.8^2) + (19)(4.5^2))/(25 + 20 - 2)}} = \frac{3.5}{4.7} = 0.74$$

也就是说，干预组成员对于环境的正面态度，依据产出测量值，比控制组对象多出了 0.74 个标准差。

发生比

项目效果大小的发生比统计量，描述在二分变量属性产出测量中的干预影响，即要测量的变量仅有两个维度，比如被逮捕和没有被逮捕、死亡和活着、充电和放电、成功和失败。这类产出测量往往以两种产出分类为基础，测量各种分类在样本、总体中所占的比例。这些数据能够以 2 * 2 列联表的形式呈现出来。

	正面结果	负面结果
干预组	p	$1 - p$
控制组	q	$1 - q$

p = 干预组中具有正面项目产出的个体所占比例；

$1 - p$ = 干预组中具有负面项目产出的个体所占比例；

q = 控制组中具有正面的项目产出的个体所占比例；

$1 - q$ = 控制组中具有负面项目产出的个体所占比例；

$p/(1 - p)$ = 干预组中个体产生正面项目效果的比例；

$q/(1 - q)$ = 控制组中个体产生正面项目效果的比例。

发生比是按照下面公式被定义的：

$$\frac{p/(1 - p)}{q/(1 - q)}$$

因此，发生比代表了干预的效果，以干预组中个体产生正面产出的比例与控制组中的相应比例相比较，以发生比值之比来表示项目效果的大小。例如，在一个“认知—行为”项目中，如果 58% 的病人在项目治疗之后，不再经历医疗意义上的精神沮丧，而控制组的这个相应比例仅为 44%，那么发生比就是：

$$\frac{0.58/0.42}{0.44/0.56} = \frac{1.38}{0.79} = 1.75$$

这就意味着,干预组中摆脱医疗意义上的精神沮丧比例是控制组相应比值的1.75倍,这就显示了项目干预对于沮丧病人的明显效果。

探明项目效果

假设一个项目实际效果的测量结果为零,但在项目干预条件下,在影响评估中并不必然得到效果为零的评估结果。在影响评估中得到的产出测量值,通常会包含一定数量的统计干扰,这些波动来自于测量误差、随机抽样误差以及将样本划分为干预组和控制组的偏差,还包括很多诸如此类的随机扰动。因此,即使产出测量过程非常规范,在干预组和控制组之间还是没有测量出实际的差异,就不可能获得用来识别项目效果的均值。同时,如果项目实际效果为零,就不能期望测量的差异会很大。另外,如果获得的差异仅来自于统计干扰,即使尽量去排除这些干扰,有时还是会让我们误将统计干扰当作项目的实际效果。总之,我们要能估计到随机波动因素的影响,这些因素会造成统计干扰。这种估计主要通过下面介绍的统计显著性检验来实现。

统计显著性

如果将实际项目效果作为项目影响评估的信号,那么,与来自统计干扰的影响就形成了一种信噪比(signal-to-noise)关系。幸运的是,统计手段为我们使用既有资料探明干扰水平提供了相应的工具。如果"信号"(在数据中观察到的项目效果估计值)与统计干扰的预期水平相比,相对较大,就可以有把握地认为:我们已经发现了项目的实际效果,而并非来自某类统计波动。另一方面,如果项目效果的估计值与统计干扰所造成的"假"的影响相比相对较小,就不能坦然地认为:我们已经观察到了项目的实际效果。

为了估计信噪比,我们必须区分作为项目效果的信号和背景性的统计干扰。关于项目效果的最优估计就是简化了的、干预组和控制组之间的产出测量值的均值差分析,在专栏10—A中可以看到经常使用到的这类效果大小统计量。如果采用合适的概率理论,我们就可以从既有数据中估计来自统计波动"假"效果的大小。这个估计值就是抽样规模(控制组和干预组中用来比较的样本数)与需要测量的产出变量分布变异程度的函数。

这种信噪比分析,一般通过统计显著性检验来实现。如果干预组与控制组之间的平均产出测量值之间差异在统计上显著,那么这个结果告诉我们,在项目假设条件下,当真实效果为零时,统计波动不可能产生与观察数据一样大的效果值。在统计检验中,一般显著水平都设定为0.05,这意味着,由随机波动导致的"假"的项目效果与观察到的项目效果完全一样的概率在5%以内。如果情况如

此,我们就有95%的置信水平来确定观察到的项目效果并不是统计波动所致。

尽管0.05的显著水平的使用极为普遍,但我们有很好的理由在具体案例中使用更高或更低的显著水平。当判断一个项目是否有效时,拥有很高的信度水平是一个非常重要的支持性因素,评估者可以设定一个更高的判断标准(门槛)。例如,显著水平为0.01,即测量结果有99%的可能性为非随机干扰造成。而在另一种情况下,比如要对计划中的干预项目进行探索性尝试,评估者也可以使用较低的显著水平——0.10,对应的置信度水平则为0.90。

不过,统计显著并不一定就意味着项目具有实际意义或重要性。统计显著的结果可能并不具有理论或实践上的重要性。有时候,只是一个不可能发生的概率统计结果而已。不过,统计显著是对有意义项目产出的最低要求,在本章的随后内容中,我们将讨论如何评估项目效果的实际意义。如果项目效果经测量在统计上不显著,按一般标准,就意味着信噪比太低,从而不能确定项目的实际作用。在这种意义上,统计显著与否就不能作为项目实际效果的唯一衡量指标。

因此,统计显著性检验首先是影响评估中评估者对测得项目效果的评价。甚至可以说是检验项目是否有效果的"有—没有"二分检验。如果观察到的项目效果在统计上显著,那么,项目效果就可能大到等于全部的产出变化。如果在统计上不显著,就不能确定产出来自于项目影响还是统计干扰,也就不具备产生科学结论所需的信度基础。

第一类错误和第二类错误

统计显著性检验为我们在项目产出数据中判断项目效果提供了一个基础和平台,但不能确保结论正确无误。随机变量有时候会造成统计扰动,这类偶然事件造成的"假"的效果有时足以影响统计显著性检验的结果,使我们对项目的实际作用进行误判。一方面,我们会误将随机扰动当作项目效果,而实际并不存在项目影响。另一方面,如果统计误差与项目效果密切相关,这些随机波动就很容易模糊掉项目的实际效果,使得即便项目效果实际上存在,在统计上也表现为不显著。以上这两类统计结论错误被称为第一类错误(type I error)和第二类错误(type Ⅱ error),参见专栏10—B更全面的讨论。

专栏10—B 第一类和第二类统计推断错误

针对干预组和控制组产出差异的统计显著性检验正确和错误的概率。

	总体环境	
样本数据的显著性检验结果	干预组和控制组的均值有差异	干预组和控制组的均值无差异
显著性差异	正确结论(概率 $=1-\beta$)	第一类错误(概率 $=\alpha$)
无显著性差异	第二类错误(概率 $=\beta$)	正确结论(概率 $=1-\alpha$)

统计功效

当然,评估者不应该设计有可能对项目效果产生错误结论的影响评估,尤其是在项目效果统计显著性这一基本结论的把握上,更应该十分谨慎。为了避免这类错误,评估者必须给予足够的注意,从而保证研究设计犯第一类和第二错误的风险较低。

犯第一类错误(项目实际上没有效果,但却在统计意义上具有显著性)的风险相对容易调控。实际上,当选定某个统计显著度 α 值以后,这一错误的最大概率就已经设定了。一般说来,如果 $\alpha = 0.05$,那么犯第一类错误的可能性也就被限定在了5%以内。

控制第二类错误(项目具有实际效果,但却无法得到统计上的显著性)则相对困难。需要规范研究计划,从而使统计检验具有更加完善的**统计功效**(Statistical power)。统计功效是一种可能性,是指对项目效果的评估具有统计显著性,并在事实上代表了给定大小上的真实项目效果。第二类错误发生的条件可能就是未获得统计显著性的补集,或者可以说,是单位1与统计功效的差值。所以,如果统计功效是0.80,那么犯第二类错误的可能性就是 $1 - 0.80$,也就是0.20。具有较高统计功效的项目影响评估设计可以展现项目效果评估的统计显著性,并且估计值大大高于某个临界值,以至于评估者无法忽略项目效果。

统计功效主要受以下几个变量的影响:①将要探讨的项目效果大小;②抽样规模;③统计显著性检验的类型;④设定用来控制第一类错误的显著水平值。一般来说,α 值会被设定为0.05,因此,如果未加注明,默认的 α 值一般都是0.05,其他三个因素需要进一步的探讨。为了确保具有完善统计功效的研究设计,评估者必须首先决定研究计划应该探明的最小效果。出于这个目的,可以将效果大小通过某个统计量比如标准化均值差(参加专栏10—A)呈现出来。例如,评估者可以选择0.20个标准差的效果量作为研究设计必须在某个统计显著水平上探讨项目效果的阈值。决定效果值的大小,要与评估者想要探明的最小规模的、有意义的项目效果相适应。我们将在后续讨论项目效果实际意义的话题中,涉及这一问题。

什么样的统计功效对于影响评估才是合适的

在选定研究用的效果“阈值(上、下限)”之后,评估者还必须决定多大的第二类错误是可以被接受的。例如,评估者可以决定不能获得统计显著性的临界值,当必须把阈值水平或更高水平的实际影响定为5%时,将使犯第二类错误的可能性也维持在0.05左右,这是第一类错误范围设定经常用到的数值。因为统计功效是犯第二类错误可能性的最小值,也就意味着评估者想要一项研究设计具有95%的统计功效来探明临界值水平或者更高水平的影响规模。类似,假如设定犯第二类错误的可能性为0.20,那么对应的统计功效则为0.80。

然后,还必须设计具有一定样本量和统计检验类型的影响评估,这样才能产

生预期水平的统计功效。抽样规模的影响相当直接，样本越多，统计功效就越高。但是，选择合适统计检验方法的作用却不那么直接，在使用统计模型时，最为重要的一个考虑就是控制变量的选择。控制变量的使用与产出测量密切相关，可以通过运用控制变量获得的项目效果，来核查项目效果的相关变异性。控制变量本身就代表了干扰因素，因而能减少统计干扰量，提升信噪比，有益于整个统计功效的提高。针对这些目的，最有用的控制变量是产出变量的前测。对于控制组和干预组之间差异的前测，可以消减将原有个体差异混入产出测量的错误。这类错误会造成与项目干预无关的产出。与项目无关的产出应该归入统计干扰中，如果其被归入干预产出，就会在分析中模糊真实的项目效果。运用经过筛选的控制变量，可以极大地提高统计功效。

为了取得良好的分析结果，控制变量必须与产出变量有较强的关系，且应纳入评定项目效果估计值的统计显著性分析中。涉及控制变量的统计分析方法包括：协方差分析、多元回归、结构方程模型和重复测量的方差分析。

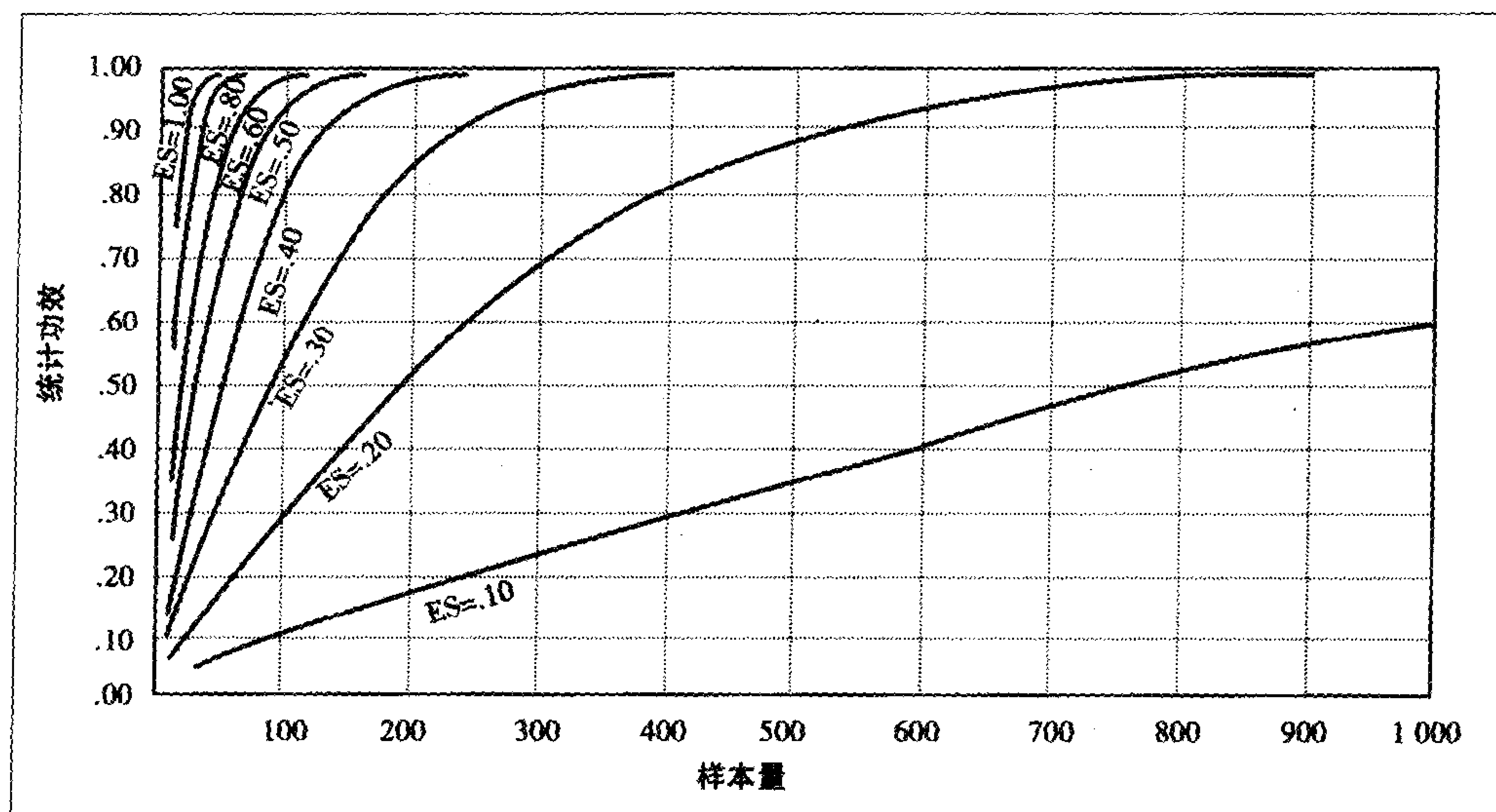
确定影响评估的统计功效是很重要的一个问题。即便评估者期望探明的项目效果很小，但却极具价值，还是要设计细致的研究规划来探明这一效果。例如，如果某项干预能够将机动车事故死亡率降低哪怕是1%，也将是很有价值的，因为能够挽救生命本身就显得弥足珍贵。相反，评估者认为要判定某些干预是有价值的，就必须依据较大的效果值；这种情况下，研究设计无力探及较小的项目效果，也就被认为无伤大雅了。譬如一个成本高昂的电脑培训计划，只有至少一半参与者获得了相应的工作职位时，才可以认为是值得推行的。总体说来，相对较高的项目效果是所有研究设计都希望获得的，也可以以较高的统计功效来探明其效果。

深入讨论统计功效估计、抽样规模、包含或不包含控制变量的统计分析，已经超出了本书范围。熟练地掌握这个领域的知识，对于顺利实施影响评估很重要。任何一个要承担这类工作的评估团队，都必须能够展示这方面的功力(如果想获得这些话题更为细致的信息，可以参见(Cohen, 1988)、(Kraemer & Thiemann, 1987)、(Lipsey, 1990; 1998)中的相关讨论)。

专栏10—C通过一个很具代表性的例子，呈现了对统计功效有较大影响的各因素之间关系。专栏罗列了就控制组和干预组均值差异的大部分统计检验(t检验或者单向的方差分析，没有控制变量，$\alpha=0.05$)而言，效果大小和抽样规模之间各种联系的统计功效。

仔细阅读专栏10—C就会发现，在影响评估中，要获得足够的统计功效是多么困难。只有当抽样规模相当大或效果大小的临界值相当大时，才能够获得相对较高的统计功效。对影响评估来讲，这两个条件经常是不现实的。

专栏 10—C　统计功效与样本量和效果大小的函数关系



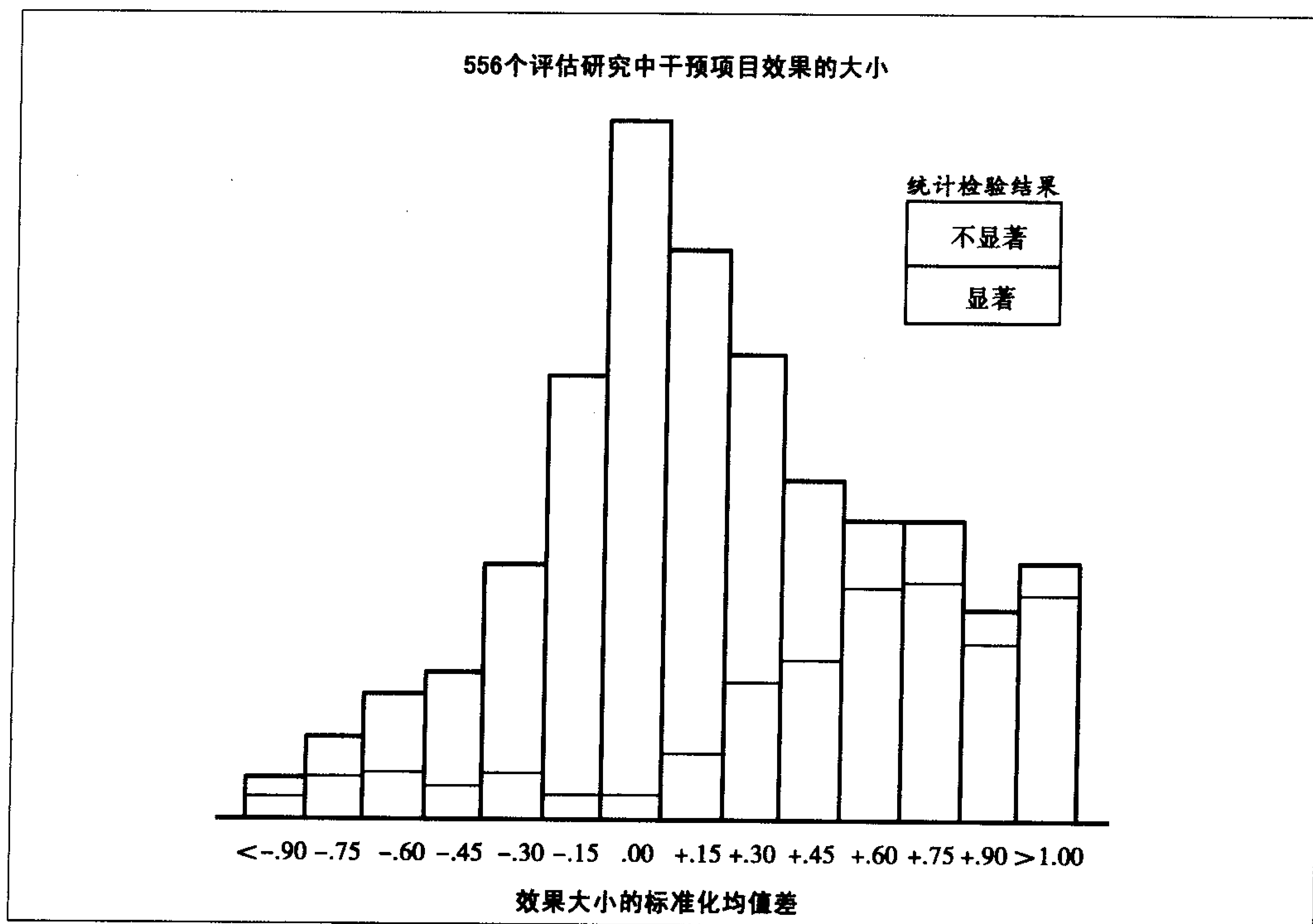
例如,假定评估者想要将犯第二类错误的风险维持在 0.05 的水平,显然与第一类错误的常用值相一致,相应的统计功效值则为 0.95。在不做深究的情况下,这不失为一个合理的目标。但是,如果项目实际上产生了有意义的效果,评估者却不能在一定统计显著水平上进行探明,就会对项目效果的真实评价造成伤害,歪曲项目效果。更进一步地说,假定评估者讨论的项目产出的统计效果是 0.20,这就是一种正面的干预效果,应该探明。专栏 10—C 的图表展现了 α 值为 0.05、没有控制变量分析的统计显著性检验,每组(干预组和控制组)样本多于 650 个,总数超过了 1 300 人。虽然可以在某些评估情景下获取这么大的样本,但与在影响评估研究中常用的样本数量相比,这个抽样规模显然过大。

如果由于实际条件限制使得样本量较小,项目效果大小的临界值就会有一个不成比例的增大,且可以轻易地探明。譬如,干预组和控制组中各 100 人,这一抽样规模在大多数评估研究工作中都很典型,只有在项目效果大于 0.50 时才可以获得 0.95 的统计功效(即第二类错误的风险为 0.05)。如果要求更小的统计效果代表重要项目任务的完成,这种评估设计在探明效果的统计显著性方面就显得很勉强。某些情景下(比如上面的情形),如果控制变量选择恰当,就可以增大项目干预的统计效果,从而允许用更小的样本量来获取相同的统计功效。

当项目效果在统计意义上并不显著时,一般说来就意味着项目没有产生预期的效果。但是,如果统计不显著是由于研究不力造成的,而不是由项目未能实施到位从而未能带来有意义的影响的话,那么对统计不显著的解释就是技术上不正确和不公正的。这些发现和结论仅仅意味着:已经观察到的项目效果并不一定比统计干扰大,干扰因素掩盖了项目效果,而非项目效果本身很小或者根本不存在。

在评估实践工作中，因影响评估技术所导致的严重问题请参见专栏 10—D。这份资料来自一项大型**事后分析**（Meta-analysis）[①]（从同一主题的多个研究中，进行项目效果的统计分析），专栏中的图表展示了效果大小的分布状态（以标准差的形式呈现，参见专栏 10—A），针对所有 556 项减少青少年违法的干预项目的效果评估的结果，都是这个分析的资料。这些评估研究分布在各种公开出版物或未出版物上，分析者试图寻找已经完成且有相应研究结果的所有合格研究，因此，极大地代表了这一特定评估研究领域的实际状况。

专栏 10—D 青少年违法行为干预项目效果的统计显著性



在专栏 10—D 中，柱形图的阴影部分表明了已有评估研究中统计检验结果显著的案例所占比例。通常而言，统计显著性是相对于最大的效果而言可接受的较高效果比例。许多项目效果虽然较小，尽管没有统计显著性，却恰恰代表了很重要的项目价值；不能呈现统计显著性是较低统计功效的一个直接结果。较大的效果估计值常常伴随较大的统计干扰量，主要是因为抽样规模过小，或未能充分运用统计分析中的控制变量。

在专栏 10—D 中，可以以 0.30 为例来讨论评估研究中项目效果的大小。在

① 事后分析，又译作元分析，是对同一主题的多项研究结果进行分析的研究，常用于多个研究结果的检验。——译者注

这些研究中,很多测量结果是以青少年违法者的重犯率(累犯率)为基础的,例如被警察连续逮捕的记录。在这样的测量中,0.30标准差的效果显示了干预组和控制组之间的差异,即在6个月的干预之后,干预组成员的累犯率为38%,而控制组成员的累犯率则为50%。这两个数值之间的比值就是干预效果,表示在项目干预之后青少年重新犯罪的数量有24%的下降(12/50)。看起来,项目的效果很有价值。尽管如此,在这一效果维度上,仍然有一半以上的项目没有统计显著性,至少评估报告是这么认为的。这些结果揭示了评估者在研究设计上的失败比没有实现的预期效果影响更大。

估计项目效果的实际意义

正如已经讨论过的,影响评估的研究设计能够得出对项目效果的统计结论;正如专栏10—A描述的那样,也允许提炼统计效果大小的类别。然而,统计效果并不必然代表项目的实际效果,统计显著水平也并不必然代表项目的实际意义。一个较小的统计效果,在实际中也可以代表相当大的项目效果。相反,一个具有较大统计效果的项目,可能具有几乎可以忽略的实际意义。例如,因患有某种疾病而必须住院的病人比重的一个很小下降,在统计上意义不大,却可能对于健康保险公司具有巨大的实际成本意义。但是,提高他们对住院照顾的满意度(统计上更大),对保险公司的**财务意义**(financial implication)则完全可以忽略。

对于项目各方和评估者来说,要解释和评价评估研究发现的项目效果,就必须首先把项目效果转化为与项目意图改善的社会状况相关。譬如,如果测量尺度具有可解释的实际意义,可以通过在原有产出测量尺度基础上重新表述统计效果大小。在青少年违法预防项目中,一个通用的产出测量指标就是参与项目一段时间后的累犯率。如果项目干预减少了24%的累犯率,就可以基于受项目影响的青少年数量和预防的违法数量进行解释。如果还熟悉青少年违法犯罪的背景,同样,可以从背景的维度解释效果的实际意义。具有内在意义的产出测量的项目效果,都比较容易解释其实际意义,比如挽救的生命数、年收入的增加量、辍学率的下降比例等。

对另一些项目效果进行解释就不那么容易。假使通过综合基本技能测试的数学部分考核,我们发现一个数学辅导项目把表现较差的6年级学生的分数从42提高到了45。我们还知道,这代表了0.3个标准差单位的统计效果,且在统计上显著。但是,在实际意义层面,这代表了多少数学技能方面的改善呢?这代表了多大的项目效果呢?很少有人熟悉这些测量总体和数学成绩测试的某个具体得分,因此大都不能直接从测试得分来解释统计效果大小。

在项目产出测量的结果并无实质意义的条件下,对统计效果的解释就要与外在参考框架相结合,将项目效果大小置于实际背景之下。例如,对于成绩测试来说,可以根据测试规则来比较项目效果。如果国家规定的数学测试标准分是

50,我们会发现数学辅导对于缩小项目参与者得分和国家标准之间的分数差很有帮助,缩小幅度达到了大约38%(分差从5分到8分不等),但是这依然无法改变他们缺乏数学知识和技能的状况。

另一个参考框架可能来自于相同学校中不同年级学生的平均成绩。假设学校6年级学生的平均成绩为47,而7年级为50,那么项目影响所造成的3分增加,就可以被当作整个年级表现水平的改善。但是,如果7年级平均成绩为67分,我们会认为3分的增长对6年级学生数学成绩的提高并没有太大的帮助。用来解释项目效果大小的一个类似框架,可以通过运用产出测量的适当"门槛"值。如果可以针对具体的产出测量设定一个合理的"成败"临界值标准,并计算出高于或者低于"门槛"的比例,这样,相应的项目效果就可以被重新转述为"成功率"的上升或下降。例如,一个应对沮丧者的精神健康项目,可以使用"贝克沮丧量表"(Beck Depression Inventory)来进行产出测量。假如项目效果的估计值为0.50个标准差,就代表了治疗组14.3的平均得分和控制组的17.8的得分之间的差异。使用这种测量工具,17~20的得分就是医学意义上沮丧的阈值。在实际中,一个很有价值的指标就是用每组测量得分在17以下的成人数来表示低于医疗临界值的沮丧水平。结果是:控制组中有42%的成员在治疗期末低于医学上的临界值水平,而治疗组中有77%的成员低于这一水平。如此,通过"门槛"值我们就轻松地评价了治疗效果,这比在任意尺度基础上的统计效果分析更为容易。

能够帮助评估者和项目各方揭示项目效果实际意义的另外一个参照基础是相类似的项目评估中的效果分布。如果我们回顾研究婚姻满意度咨询项目的相关影响评估文献,对这类项目的效果进行一个事后分析,发现平均效果大小为0.46左右,大部分项目效果处在0.12到0.80之间。就手头的资料而言,假若评估者在某个婚姻项目的影响评估中发现,项目干预对婚姻满意的影响为0.34,我们就能够判断这个项目的影响水平等级为中上。当然,如果项目服务于某个极为特殊的对象群体或者在一种极其困难的环境中实施,我们也有很好的理由理解处于平均绩效水平之下的项目。然而,对类似项目的效果进行普遍比较的方法,如果加以利用,就能作为评价项目效果大小的工具。

没有十全十美的方法用来解释项目效果的大小并评价其实际意义,但是,我们已经论及的这些方法通常都是有价值且可以利用的。显然,统计意义的简单陈述或统计效果大小的判定极少能很充分地描述项目的真正效果。因此,评估者应该学会一种或更多种将统计效果转化为实际效果的方法,进而在项目具体操作的实践背景中,更为贴切地解释项目效果。特定类型和在特定背景中最有意义的解释框架,会视具体情况而有所不同,评估者需要更为灵活地发展出适用于具体项目的解释框架。在专栏10—E中,我们对讨论过的方法和在某些条件下可能有用的方法,都进行了系统说明;但还是应该承认,这份清单并不能穷尽所有的可能性。

专栏10—E 实际应用中用来描述统计效果大小的一些方法

原始测量尺度上的差异:当原始测量数据具有实际意义时,项目效果可以直接通过控制组和干预组之间结果差异的比较来表述。例如,一个预防项目实施之后健康服务的货币价值,或意图减少放射时间的项目实施之后的住院天数,在各自的背景框架之下都具有实际意义。

以检验规范为基础的比较或总体绩效水平比较:对于意在改善目标人群相关状况使其达到主流水平的项目来说,项目效果可以表述为减少干预对象与主流群体之间的相关差距。例如,针对不能顺畅阅读的儿童开展的项目干预的效果,可以通过分析项目实施之后,受干预儿童阅读水平与正常儿童之间的接近程度来衡量。所谓正常水平可能来自公开的检验规范,也可能来自和项目参与者同级、同校的其他儿童的总体平均水平。

标准组之间的差异:在项目实际背景下,当项目群体在产出测量上具有可识别的标准差异时,各自产出测量上的差异就可以被认为是项目效果。假若一个精神健康项目可以很正式地使用沮丧量表(depression scale)来识别病人在引入治疗之后的差别,这些病人有的按照门诊病人进行照顾,而有的按照住院病人进行治疗,通过沮丧程度量表的测量就可以获得项目效果,从而可以通过对门诊病人和住院病人得分上的比较,来展示其与已经发现的结果差异在多大程度上相关。

治疗比或其他标示成功的"临界值":当产出测量的某个值可以被设定为评判项目成功与否的临界值(阈值)时,具有成功结果的干预组内部的样本比例就可以与控制组具有类似结果的比例相比较,揭示项目效果。例如,一个职业项目对工资收入的影响就可以通过干预组中收入在联邦贫困线以上的家庭比例与控制组中相应比例的对比来呈现。

"成功临界值"之上的样本比例:要指明项目的效果,成功率是一个很有用的概念,即使判定成功与否的临界值具有任意性,也可以通过"成功率"来描述项目的实际意义。例如,控制组的平均结果值可以被设定为"门槛"性的临界值,在一般情况下,控制组50%的对象将在均值之上。处于临界值上的干预群体样本比例就可以表示项目效果的大小,譬如如果55%的干预对象处于控制组结果均值之上,那么与项目实施使75%的样本处于均值之上的情况相比,项目效果就比较小。

类似项目之间的效果比较:可以根据已有的类似评估研究积累的结果,为识别项目的统计效果提供有用信息。通过与其他类似项目的比较,来识别项目效果的大小。事后分析可以系统地整理和报告这种类似项目的干预效果情况,对于这种识别活动尤其有益。因此,一个戒烟项目实施之后,如果根据连续不抽烟天数,其项目效果标准差达到0.22,就可以被认为具有较大的实际项目效果。如果其他类似项目的产出测量值的平均效果仅为0.10左右,而自己的相应数据为0.50,就可以认为这个项目具有比较大的实际项目效果。

传统指引:科恩(Cohen,1988)曾将社会科学研究中的影响效果划分为"大"、"中等"和"小"三等,并提出了一般性的指导原则。尽管其是在统计功效分析的背景下提出的,但至今仍然是判断干预效果大小的、广泛应用的经验法则。例如,对于效果大小的标准差,科恩指出,低于0.20是小规模影响,而0.50是中等规模影响,达到或超过0.80则算作是大规模影响。

分析项目效果的差异性

直到现在,我们对于项目效果的关注主要是,在与相应控制组比较的基础上干预组的总体平均效果。然而,项目效果对干预对象各个次级群体以及所有项目产出的作用很少是完全相同的,项目效果的差异性分布对于评估研究也非常重要。除了相关利益群体的产出测量,除了干预组和控制组织之间关系的分析,我们还需要将其他变量纳入到对项目效果差异性的测量之中。当关注目标人群中各个次级群体在项目效果上的差异时,可以使用调节变量定义次级群体,从而进行分解分析。同时,为了核查某个产出变量的变化如何影响到另外一个产出变量的变动,在分析中,就必须在所有产出变量中找到一个潜在的调节变量来进行检验分析,从而弄清楚项目作用的因果链条。接下来的部分,我们将描述项目效果的差异性和变动性如何能够与调节变量或中介变量相关,并分析评估者如何才能够揭示这些关系,从而确保能够更好地理解项目效果的性质。

调节变量

调节变量(Moderator variable)可以在项目影响评估中用来界定次级群体,描述影响评估中次级群体的特征和差异性。例如,当考虑项目效果对男性、女性参与者是否不同时,性别就可以作为一个调节变量。为了探明这种可能特征,评估者将把干预组和控制组成员划分为男性和女性次级群体,将平均项目效果分解到性别,从而比较男性与女性在项目效果上的差异。评估者可能发现,项目效果对女性更大(更小),而男性则相对较小(较大),并且这种差异具有统计显著性。

通过调节变量进行分析,可以揭示目标人群在次级群体划分基础上的项目效果差异,这是相当普遍的做法。主要的人口变量,比如性别、年龄、民族和社会经济地位,经常被用来作为调节变量,用来分析具体社会项目在效果上的群体差异性。很显然,除了项目的总体平均效果之外,让项目方清楚干预对于哪些特定群体更为有效或者无效也是很有帮助的。因此,调节变量的识别和分析是影响评估的一个重要方面。通过这种分析,可以识别出项目影响对于次级群体的效果大小,识别项目效果的敏感群体或盲点群体。对于所有干预对象来说,即便项目效果的总体平均水平很低,这种方法也能深度挖掘项目产出数据,从而加强项目绩效的整体结论。例如,关注从项目中获益最少的群体,从而探寻针对这些群体而言更加行之有效的干预方法,将是加强项目效果从而提高整体绩效的快捷方法。

如果评估者在影响评估开始时就定义了不同的次级群体,就可以大胆而清晰地探明项目效果在次级群体上的差异性。这种情况下,不存在选择性偏差。例如,干预对象显然不会因为项目的选择过程而改变性别。然而,假如在项目干预阶段才定义次级群体,那么选择性偏差就会起作用。如果干预组或控制组的

某些成员在被分配到干预状态或控制状态之后脱离了原来的状态,那么影响其行为的任何力量都可能影响到项目产出。相应地,我们对于项目效果的分析就需要考虑这些次级群体带来的选择性偏差。

如果评估者已经测量了相关调节变量,那么,对于分析项目效果在干预人群中的分布,将尤其有价值。当把最需要项目干预的人群纳入到影响研究中时,经常会发现项目干预对他们的效果最小。譬如,一个职业培训项目,将对有工作经验和相关工作技能者的职业发展产生更明显的效果,而几乎没有工作经历或工作技能的长期失业人员则较少受益,尽管他们才是最需要服务的目标人群。虽然这个结果本身并不令人惊讶或必然错误,但调节变量分析却能够揭示最需要服务的个体是否从根本上获益。如果项目的积极影响对最需要的目标群体是不必要的、无效的或微不足道的,那么,与那些能为最需要者带来哪怕是很微弱的项目收益的情形相比,评估的实施和项目的改进都将是极为不同的。

在这个例子中,职业培训项目的效果差异如此巨大,以至于项目的总体平均效果显示:尽管项目没有对长期失业的次级群体产生影响,但他们后来的工资还是比较高。如果没有调节变量分析,总体正面影响将会掩盖这样的事实——项目对于某个重要群体来说是无效的。这种对实施效果的遮蔽也有别的方式。有时,项目总体效果可能是可以忽略的,也就意味着项目是无效的;然而,调节变量分析却可以揭示项目对个别次级群体的显著影响,即在总体平均水平上被抹杀掉的影响。这种情况会经常发生。譬如提供普遍服务的项目,会覆盖到项目意图之外的很多个体。例如,在中学普遍实施的毒品干预项目,就会涉及很多并不使用毒品的学生,尽管他们极少有可能使用毒品。无论项目本身多好,都不能对这些学生的毒品使用结果产生显著影响。因此,项目影响检验的一个重要程序就是调节变量分析,可以核查高危人群的项目实施结果。

因此,调节变量分析的一个重要角色就是避免对项目绩效下不成熟的结论,因为仅仅建立在对项目效果整体平均水平概括之上的项目绩效判定有其局限性。一个影响全面且积极的项目对各类参与者并不一定都是有效的。同样,一个没有显示出全面影响的项目对一些次级群体可能是非常有效的。另外一种可能性,就是积极和消极影响并存,这种情况虽然很少,但却对调节变量的诊断特别重要。一个项目也许对其中某个参与群体有系统的、消极的影响,但在整个影响评估中只是一个方面,因为对另一些次级群体的影响可能是积极的。例如,在不良行为青少年群体中实施的项目,可能会成功地降低严重违法者的不良行为,但对违法行为不是很严重的参与者来说,在项目实施中,他们的行为可能会被有严重犯罪行为的同伴群体影响,实际上还可能增加他们的犯罪率(cf. Dishion, McCord, & Poulin, 1999)。由于严重违法者的比例高低不同,这一否定性的效果,在对整个项目中违法行为的总体平均效果的影响上可能并不明显。

除了揭示项目效果上的差异,评估者还可以运用调节变量分析来检验假设或预期,以便核实应该呈现什么样的差异性效果。这对于分析影响评估结果的一致性尤其有帮助,还有益于加强从研究结果中所总结出来的、关于项目效果的

整体结论。第7章、第8章和第9章讨论了影响评估中偏差或模棱两可的多种来源,这些偏差或含糊性使我们对项目效果的有效评估变得更为复杂,有选择性地分析项目效果差异的类型,提供了另一种核查项目合理性的方法——通过这种方法,来确证是否由项目自身而非其他未加控制的影响因素,产生了我们观察到的项目效果。

例如,一种有用的研究方法是"剂量—反应"分析(Dose response analysis)。这个概念源自医学药物研究,基本假设是:当所有其他条件保持不变的情况下,更大剂量的药物治疗应该能产生更好的治疗效果,至少会达到最佳剂量水平的治疗效果。当然,保持所有条件不变是极其困难的,但是一般来说,只要数量、质量或者是服务类型上存在一定的差异,这种方法对于要实施调节变量分析的评估者就仍然是可供借鉴的。例如,假设一个项目设有两个服务提供站,为类似客户提供服务,如果项目在其中的一个服务点被更好地得到了实施,那么按照预期,项目效果在这个服务点就相应会更为突出。但是,如果评估结果不是这样,尤其是当项目效果在实施比较薄弱的服务点却更为突出时,这种不一致的情况就会促使我们对先前的假设提出疑问:测量得到的项目效果到底来自于项目本身,还是其他来源。当然,也可能有一些对这种不一致性的合理解释:譬如搜集的实施资料不准确,或是干预和控制对象的原有差异未加识别。但是,调节变量分析依然有潜力促使评估者意识到支持项目影响评估结论的逻辑框架的一些问题。

专栏9—I给出了一个具有类似特征的例子,描述了一个经典的评估项目,即针对酒精测试对英国交通事故效果的时间序列评估。因为评估中使用的时间序列设计对于分离项目效果不是特别有说服力,因此项目评估的一个重要组成部分就是进行调节变量分析,通过分析来核查周末晚上时段和工作日交通高峰时段在项目效果上的差异。研究者的预期是:如果项目实施的确在可观察的交通事故减少上起作用,那么项目效果在容易出现酒后驾车的周末夜晚比往返上班的时间段更为显著。结果证明了研究者的这种预设,支持了以下这种结论:项目实施是有效的。但如果核查的结果显示往返上班时段的项目效果比周末夜晚时段更为突出,那么项目实施的有效性就会大为削弱。

这种意在探索项目在达成可测量效果中作用的调节变量,其分析逻辑就是核实项目差异性效果。评估者认为,如果项目依照预期设想实施,而且确实产生了特定的效果,某些情况下的效果应该比其他情况下的效果更大。例如,要改变的目标行为已经极为流行,或提供更好的服务,再或者目标群体对将要发生的反应更为适应,等等。如果恰当的调节变量分析证实了这些预期,就会加强对项目效果的肯定。显然,如果这些分析不能证实先前的预设,就会使评估者有理由相信,除了项目作用之外,还存在其他一些影响项目评估结果的因素。

一方面,我们意识到了调节变量分析对于分析项目效果结论合理性的价值;另一方面,我们也必须指出其局限性。例如,项目目标人群所受到的干预不是随机的,所以,如果基于与服务数量相关的调节变量进行群体之间的比较分析,就

可能产生偏差。例如,收到最多干预的参与者可能是问题最为严重的一些人,这种情况下,简单的“剂量—反应”分析将会展现相对其收到的服务而言小得多的项目效果。然而,如果评估者分析具有相同问题程度的群体成员,“剂量—反应”关系就会起作用。很显然,调节变量分析对项目效果检验的可解释性有诸多局限,这也是我们把它仅仅作为影响评估设计补充方法而非对等替代性方法的原因。

中介变量

在评估中,能够唤起人们注意的、事关项目效果差异性的另一个方面,就是产出变量之间可能存在的中介变量关系。如果存在中介变量,中介变量就是项目效果的最直接结果,其变化进一步影响下一个长远的项目产出。所以,中介变量就是一个中间变量,处于项目实施和一些核心产出之间,是因果关系路径上的一个中间环节,通过它,项目实施带来了最终结果上的变化。正如在第 5 章和第 7 章讨论到的,在项目影响理论中被识别的最近产出,都可以被称为中介变量。

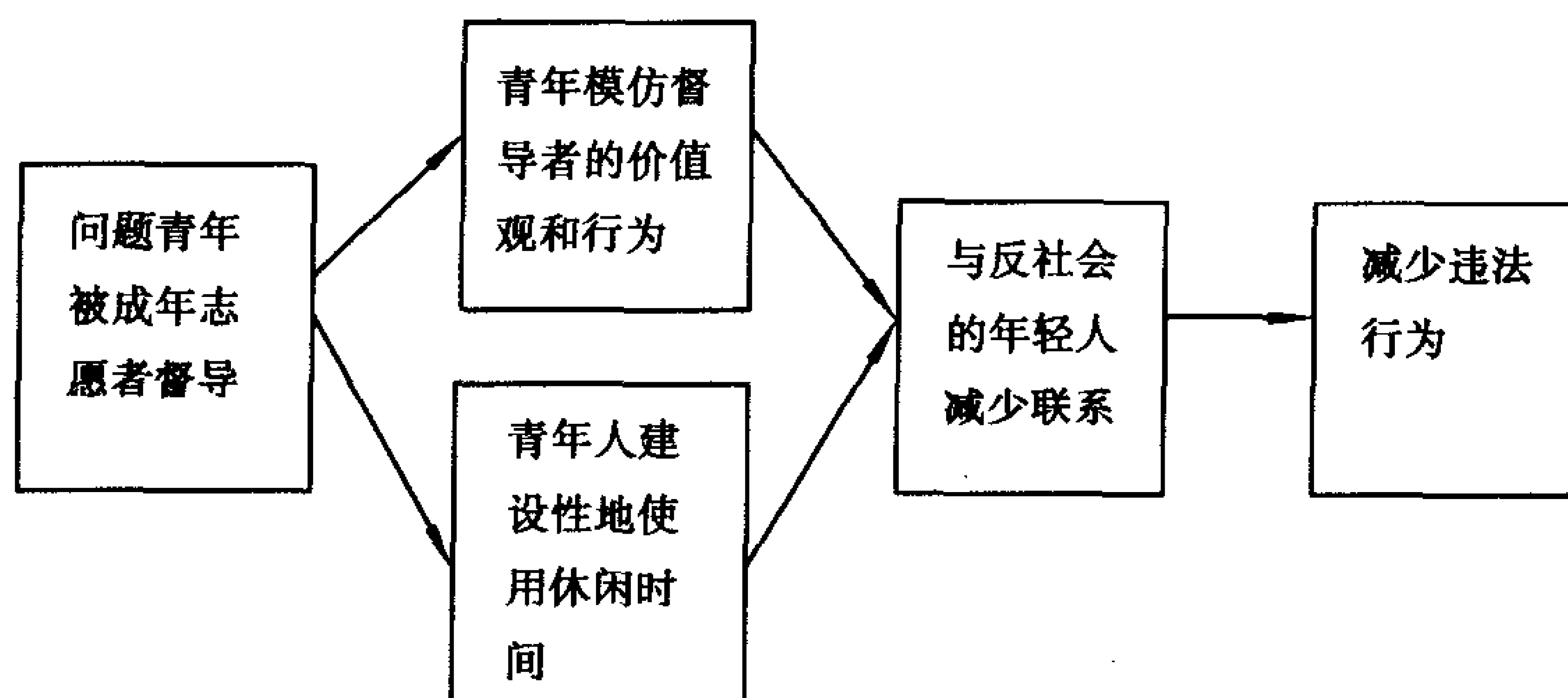
和调节变量一样,中介变量的存在只有在以下两个原因存在的条件下才有意义。首先,通过对中介关系的探究,有助于评估者和项目方更好地理解项目实施的产出,即在项目对象中究竟发生了什么样的变化。从而促进人们对改善项目的推进方法作进一步思考并调整项目,从而达到更好的效果。第二,检验项目逻辑结构中变量之间的中介关系假设,这是另外一种探明评估结果的方法。假如项目有一个预设的效果目标,我们就可以由此来判断,项目产出是否与预期目标完全符合。

在第 7 章论及的项目影响理论中,我们可以通过其中的一个案例来对中介关系分析进行阐释,即专栏 10—F。专栏示意图展示了一个项目的关系结构,这个项目通过成年志愿者对问题青年进行督导,从而达到干预效果。这个项目预期产生的最终结果是对问题青年的违法行为进行干预。隐藏在影响理论框架背后的因果关系路径,就是通过与督导者的互动来影响问题少年,使他们模仿督导者积极的价值观和行为模式,并引导他们更合理地利用闲暇时间。相应地,这个项目希望通过以上措施来减少问题少年与反社会同辈群体的接触,从而最终减少违法行为。在这个假设关系路径中,正面的价值观和行为以及建设性的闲暇时间利用是项目效果与同辈群体之间接触的中介变量。同样,也假定与同辈群体接触是价值观、行为变化、闲暇时间利用与减少违法行为之间的中间变量。

简单而言,我们暂时只把正面价值观假定为中介变量,在督导者对违法青少年群体成员之间的互动干预中,起着中介作用。要检验这些变量之间是否存在中介关系,首先就要确定项目效果具有最终产出(与反社会同辈群体的接触)和近期产出两个层面(价值观)。如果近期(直接)产出并不受项目影响,那么就不能作为中介变量起相应作用。同理,如果最终产出并未显示项目效果,也就没什么中介变量可言了。如果两种效果都存在,那么采用统计性控制变量手段,对中介关系的检验就能帮助弄清楚近期产出和长远产出的关系。进行这种检验的一

种可行方法就是包含控制变量的多元回归分析(第9章中已讨论过)。就手头的案例来说,与反社会同辈群体的接触在分析过程中将是一个因变量。而解释变量则是正面价值观和群体地位(一个二分变量:1 = 干预组,0 = 控制组)。如果正面价值观与反社会同辈群体之间有显著关系,当在统计上控制了督导项目中的参与之后,中介关系就呈现出来了。

专栏10—1 近期产出与长远产出:以一个项目影响理论为例



中介变量关系检验统计方法的细节,可以参见巴里、肯尼(Baron and Kenney, 1986)和迈金尼与德威尔(MacKinnon and Dwyer, 1993)的有关著作。就当前目的而言,我们主要关心能够从这些分析中获得什么。假如在所有针对督导项目的影响理论中,我们对中介关系都进行分析,那么,毋庸置疑,将可以从这些分析中获益:有利于增进那些旨在达到一定效果的项目合理性。而且,即便中介变量的分析与预期并不一致,对于项目仍然具有诊断价值。例如,假如年轻人的正面价值观有项目效果,但却没有显示其与反社会年轻人互动或与违法行为相关的中介变量关系;而同时,闲暇时间的使用却显示变量之间的强中介变量关系。那么,上面这种类型的结果就意味着,尽管青年人的正面价值观受其和督导者关系的影响,但并不导向进一步有意义的项目产出。因此,如果这个项目培训的督导者投入更多的精力帮助青年人建设性地使用闲暇时间的话,就会在消减反社会和违法行为上有更大的项目进展。

事后分析的角色

通过以上章节和本章对影响评估的讨论,我们已经强调了使用项目影响评估结果的好处,影响评估在项目设计和分析新的项目影响评估时已经得到了很广泛的应用。另外,如果每一新的项目评估都建立在已有研究发现之上,那么关于项目干预的知识也会不断增长。

如果我们对单个的影响评估报告进行仔细的研究,并对专业期刊和互联网

上的有关研究成果进行细致的回顾,就可以从中获益很多。尽管评估者的具体目的不同,但最为有益的一种总结就是增加事后分析的数量,从统计上对评估结果进行综合,或对已有的成千上百的影响评估进行总结。不幸的是,在某些情景下这类综合研究并不能应用,因为对同一级别或同一类型的影响评估数量不大的事后分析,常常不够到位。同时,对于很多干预项目,尤其是大规模的项目,并不一定有足够数量的类似评估可供事后分析。即使真的进行了事后分析,也会由于已有研究数量不够,而不能为评估者提供有价值的信息。

不过,对于评估者而言,理解事后分析的实施方法是很有用的。在一个很典型的事后分析中,首先要搜集特定干预活动或某类社会项目的可用资料。然后,在已选项目产出的基础上,对项目效果进行编码,按照分类的统计量对效果大小进行编码(分类请参见专栏 10—A)。事后分析还涉及对项目方法、项目参与者和干预性质的各种描述信息进行编码。然后,将这些信息汇编成可用不同方法分析的数据库,依据项目效果的变化和与差异相关的各类因素,对不同评估项目的不同结果进行统计分析(Cooper and Hedges, 1994; Lipsey and Wilson, 2001)。这些分析结果对设计与事后分析中相类似的影响评估方案,很有借鉴意义。另外,通过综合不同评估者对特定类型项目效果的众多研究,事后分析对于促进相应领域的评估发展也很有意义。下面就对这些贡献分别进行讨论。

了解一个具体的影响评估

任何引导和总结大致相同类型干预的事后分析,一般来说,都可以为研究设计提供有用的信息。相应地,评估者必须特别注意,需要将事后分析工作视作是相关研究(应该进行影响评估)的一般性评述。专栏 10—G 是对干预攻击性行为的学校项目的一个事后分析结果,这个分析得到的信息是很有用的。

事后分析主要关注由干预产生的统计效果大小,因此可以提供影响评估的信息。事后分析必须考虑恰当的统计功效,例如,评估者必须对项目可能产生的效果大小有一个大致的了解,必须明白什么样的最小效果才是值得探讨的。显然,事后分析提供具体项目领域的总体平均效果水平,也涉及不同项目之间的差异性分析。根据分析得到的效果均值的标准差,让评估者加深对项目效果分布和作用幅度的认识,从而对要评估项目的最小效果和最大效果做出估计。

当然,项目效果大小的意义对于不同结果也是不相同的。许多事后分析考察可利用的评估研究中产出变量的不同范畴。这种信息使评估者了解其他研究的预期效果和研究发现。当然,事后分析有时候也会用处不大,假如项目评估仅仅关注具体项目的结果,而不考虑其他类似项目评估的结论,事后分析就会失去其本该具有的意义。然而,即使这样,相似变量的事后分析结果,譬如态度、行为、绩效,等等,也会有助于评估者同时考虑项目产生预期效果的可能性和效果的预期大小。

同样,在完成了一个项目的影响评估后,评估者就能够利用相应的事后分析方法,对已经在评估中发现的项目效果大小进行评价。通过同类项目范围内影

响评估的事后分析，可以建立项目效果大小的数据资料库，这个数据库构成了一个描述模型，既包括对典型项目效果的描述，也包括效果变动的范围。每个评估者都可以利用这些信息进行基本估量，评估在已有项目中已经进行过评估的各种效果是否典型地代表了某一类项目。当然，这种评价模式必须考虑各方面的差异，包括手头项目的结果和事后分析结果之间在人群特征、客户和环境等方面的不同。

事后分析系统地展现了项目特征和项目效果在不同产出结果之间的关系，这不仅易于评估者比较各个项目的效果，也给评估者提供了一些评估线索，即什么特征的项目能最有力地获得干预效果。事后分析在专栏 10—G 中进行了概要性的阐述：比如在事后分析中，如果项目由一般人（父母，志愿者）执行，就比由老师执行效率更低，一对一的模式会比团体模式产生更好的结果。因此，评估者在对一个学校干预攻击性行为的项目影响评估进行指导时，可能会把注意力放在项目的人群特征上。

专栏 10—G 事后分析：以学校干预攻击性行为的项目效果为例

许多学校都实施意在预防或减少攻击性和破坏性行为的项目。为了调查这些项目的效果，就从对这些项目进行的影响评估中抽取了 221 个，对其评估发现进行事后分析。

首先，全面搜集这类学校项目的出版物和未公开出版研究报告，发现这些项目在一个和多个年级实施，涉及从学前班到高三的每个年级。如果想把每一项研究纳入事后分析，就必须报告其对攻击性行为的产出测量（比如打架、欺负弱小、个人犯罪、行为问题、行为失范、行为出格等），还必须满足一定的方法论标准。

然后，对每一项研究中项目干预攻击性行为效果的标准化均值差进行计算。最为普遍的项目平均效果大小如下。另外，对效果大小的调节变量关系分析显示，四种情况下项目效果更大。

治疗或咨询服务（治疗服务，比如群体或者个体咨询和个案管理）	.33
社会能力培训、认知行为培养（通过培训开发社会技能、理解和控制愤怒的感情、解决冲突，还包括其他一些使用认知行为方法的内容）	.27
行为管理技术和教室管理技术（使用各种行为技术，比如报酬、符号经济、权变契约，从而形塑行为）	.22
社会能力培训，但不是认知行为培训（开发社会技能，达成理解和控制愤怒，解决冲突，使用指导性的方法而非认知性的行为技术）	.20
多模型社会项目（对以下内容中的至少三种能力进行干预：社会能力培养、咨询、教室管理、父母培训和学术服务等）	.15
• 问题严重的儿童是项目对象 • 项目得到了很好实施 • 项目由老师们进行管理 • 使用一对一的、个体化的项目形式	项目效果更大

资料来源：Sandra J. Wilson, Mark W. Lipsey, and James H. Derzon, “The effects of School-Based

Intervention Programs on Aggressive Behavior: A Meta-Analysis." *Journal of Consulting and Clinical Psychology*, 2003, 71 (1): 136-149. Reprinted with permission from the American Psychological Association.

了解项目评估领域

除了支持具体项目的评估以外,事后分析在评估领域的另外一个重要功能,就是概括地评估已实施项目的普遍特征。尽管在某些方面每一个项目都是独一无二的,但这并不意味着在评估结果中不应该去争取发现某些共有的特征。进行事后分析可以使我们对项目如何起作用、项目对象是谁以及项目实施的环境加深理解。这方面的可靠知识不仅有利于评估者完善其实施项目评估的关注点和研究设计,而且能够为决策者提供知识平台,使他们明白什么样的方法对于解决特定社会问题是最好的。

事后分析已经成为综合评估者和其他研究者对社会干预项目效果研究的一个基本方法。当然,对很多研究进行概括有一定难度,因为每个社会项目本身就很复杂,每个项目产生的效果也有不少差异。但是,在不少项目领域都有相对固定的项目进程,从而可以大致评判干预模型的有效性、不同干预的项目效果性质和大小,还可以找出确保项目成功的主要因素。另外一个好处就是可以认识和分析影响评估中方法的角色,方法总会影响评估结果。

对于评估者来说,综合多个影响评估结果的不断努力有一个很重要的应用:必须全面地报告每一个影响评估的研究结果,进而使其进入事后分析。在这一意义上,评估领域本身成了每一个具体评估项目的项目方。就像所有其他项目方一样,评估领域为评估者提供了供其纳入考虑的明确信息,当他们设计和报告评估项目时,就可以进行参考。

小　结

- 影响评估探明项目效果及重要性的能力在很大程度上依赖于项目效果本身的规模和大小。因此,评估者必须对项目效果的统计大小和实际大小进行区分,在对其进行描述时,必须清楚两者之间的关系。
- 为描述项目统计意义上的效果大小,评估者经常使用诸如标准化均值差、发生比(结果是二分的)之类的统计量来进行测量。
- 在存在一定统计干扰的情况下,在评估中往往很难探知较小的统计效果。**信噪比**(Signal-noise)最初用来判断项目的统计效果是否大到足够与背景干扰相区分,并通过统计进行显著性检验。
- 当尝试从统计上来探明项目效果时,评估者也可能从结果资料中得到错误的结论。一个统计意义上影响显著的项目,可能实际上并没有什么效果(第一类错误);另一种情况则是,统计意义上不显著的项目效果,可能实际上却造成了真实的影响(第二类错误)。
- 为了防止针对统计效果的错误结论,评估者必须小心谨慎,以确保研究设计犯第一类错误和第二类错误的可能性比较小。一般来说,第一类错误统计检验的显著性水平限定在 $\alpha=0.05$。但是,防止或减少犯第二类错误的难度很大,需要研究设计具有较高的统计功效。

- 要探明某个具体项目在最低显著度限定下的效果，研究设计需要一定的统计功效。统计功效受到干预组与控制组样本量的影响，还受到所选用的统计显著性检验方法的影响，尤其受到统计检验中涉及的控制变量的影响。
- 尽管探知项目效果的能力依赖于项目效果统计意义上的大小，但较大的统计效果并不一定就意味着项目影响具有实际重要性。我们有必要解释统计效果所对应的实际重要性，这就需要把统计效果转换为与项目准备改善的社会状况相关联的概念。没有通用的方法可以完成这种转换，但评估者自己可以有很多选择来完成这一工作。
- 尽管有一个总体意义上的项目平均效果，但对不同次级群体，项目效果通常并不一样。调查调节变量是项目影响评估的一个重要方面，这些变量使不同人群之间表现出明显差异。这种分析方法可以解释项目效果对某些特定群体是特别大或者特别小的，还允许评估者进一步挖掘项目产出资料，从而完善和加强项目绩效方面的结论。
- 中介变量分析则可以探索最终项目效果和较近的项目效果之间的关系，并分析两个阶段效果上的变化，从而进一步依据项目影响理论，分析一个变量对另一个变量的因果关系。这些联系界定了中介关系，可以使评估者和项目各方对发生在目标对象中的变化过程加深了解，这种变化是项目影响下的结果。
- 事后分析的结果可以为评估者设计影响评估提供信息和知识支持。事后分析结果可以显示同类项目干预影响下的项目产出在效果大小和作用幅度方面的诸多不同。这就为类似项目的干预效果之间在影响评估基础上进行比较，提供一个交流平台。
- 另外，事后分析已经成为一种基本的综合分析方法，为评估者和其他研究者交流社会干预的效果研究提供了平台。就这个角色而言，事后分析扩展了评估领域的知识，使我们对过去一些年内的众多影响评估能有一个总体性的把握。

基本概念

效果大小统计量(Effect size statistic)：估计项目效果大小的统计量。以标准化的形式显示项目效果的大小，从而使不同变量或度量标准之间可以相互比较，两个常用的效果大小统计量是标准化的均值差和发生比。

中介变量(Mediator variable)：在影响评估中，一个可能的结果是项目操作中某些因素影响干预对象，进而影响到后期的项目产出。因此，中介变量就是在干预变量和项目产出之间的干预因素，在干预变量和项目产出之间建立了因果联系。

事后分析(Meta-analysis)：一种分析影响评估的统计方法，通过类似或相同的多个项目干预之间的比较，以可量化的结果总结和比较一组类似研究发现。

调节变量(Moderator variable)：在影响评估中，作为一个变量，比如说性别或者年龄，使得项目效果根据其划分的次级群体呈现出差异。

发生比(Odds ratio)：表示项目作用大小的统计量，通过干预组和控制组对比，分析在“成败”结果概率上的比值。

标准化均值差(Standardized mean difference)：表示项目效果差异的统计量，用来描述干预组和控制组按照标准偏差单位，在平均值上的差异性。

统计功效(Statistical power):一个观察到的项目影响,当事实上代表了一种真实的项目作用时,这种作用在统计意义上显著的可能性。如果某一真实的项目效果并未被发现在统计上显著,即为犯第二类错误的结果。因此,统计功效是单位1与犯第二类错误概率的差值(参见第二类错误)。

第一类错误(Type Ⅰ error):统计结论错误的一种,在统计意义上项目效果是显著的,但事实上,项目却并未对项目对象产生效果。

第二类错误(Type Ⅱ error):统计结论错误的另一种,在统计意义上项目的影响是不显著的,但事实上,项目的确对项目对象产生了效果。

效率测量 

项目成功执行的程度如何、获得的产出如何，对项目经理、项目各方和决策者而言，都是不可缺少的信息。然而，几乎在所有情况下，最重要的是要知道与项目成本相比较，项目的产出如何。事实上，项目是否给人留下深刻影响，正如大多数日常生活判断或正式决策过程一样，在决定是否应该扩展、继续或结束项目时，比较项目成本与收益是最重要的考虑。

效率评估（成本—收益和成本—绩效分析）为项目成本与效果之间的联系提供了一个分析框架。除了提供确定资源配置所需的信息以外，效率评估通常还能帮助项目得到来自规划团体和政治实体方面的支持，而后者掌握着社会项目的命运。

两种分析程序的技术性都很强，所以本章只能进行简要介绍。然而，因为在所有影响评估中，为达到预期变化所要付出的努力或成本都不是很清楚，所以评估者必须理解效率评估的思想，尽管不一定具有进行这些技术分析所需的训练和能力。

在社会干预的决策中经常提到效率问题,正如下面这些例子所展示的那样:

- 政策决策者必须确定如何在不同教育项目中配置资金,譬如,从为新移民安排基础文化教育到为失业工人提供假期培训教育。从已经完成的评估中,所有人都能了解到项目带来的实质影响。这其中,政策制订者的一个重要考虑,将是弄清楚项目**收益**(Benefit)(积极的产出,包括直接的和间接的影响)是否超过了所需**成本**(Cost)(产生干预效果所必需的直接和间接的投入)。
- 政府部门考察目前正在进行的多个国家级疾病防治项目。如果有额外经费要投入到控制疾病项目中,哪个项目的投资回报率最高?
- 在司法领域,已设立了许多项目来帮助减少犯罪。对于司法部门来讲,就成本—绩效而言,哪个项目最有效?如果有了改善目前成本模式的政策,如何才能获得最佳效果呢?
- 如果私营基金组织成员正在讨论是否应该为已婚妇女发起一个为家庭建设和工作培训而设的低息贷款项目,从而帮助她们提高家庭收入。他们的决定应该如何?

这些就是所有规划者、基金组织和政策决策者经常碰到的、有关资源配置难题的几个例子。决策者必须一遍又一遍地决定如何将有限的资源进行配置,并使其作用发挥至极致。即使是在最幸运的情况下,许多项目的探索性规划也必须展示其在获得预期净影响方面的效果。所以,在确定资助规模较大的项目时,就必须考虑每个项目的成本与收益关系。尽管还有其他因素的影响(包括政治和价值因素),在一定花费水平下,选中的项目往往是对最合适目标人群起作用最大的一个。这种简单的概念就是成本—收益和成本—绩效分析(一种系统地进行资源配置的技术)。

成本—收益和成本—绩效分析都是被用来判断项目效率的方法。正如我们将精心描述的一样,这两种分析的不同之处就是表达项目产出的方法不同。在成本—收益分析中,项目产出是用货币形式来表示的;在成本—绩效分析中,产出是用实质性效果来表示的。例如,减少吸烟人口项目的成本—收益分析关注的是反吸烟项目的开销和因吸烟人口减少导致治疗相关疾病开销降低之间的差值。而成本—绩效分析所考虑的应该是,将每个吸烟者转化为非吸烟者所需花费的开销是多少(在本章的以下部分,我们将讨论如何确定使用成本—收益或者成本—绩效分析)。

资源配置分析最基本的程序和概念是20世纪30年代在为公共投资制订决策标准的过程中渐渐形成的。其早期在美国的应用,主要是在水资源开发利用方面;在英国,则主要是在交通运输投资方面。第二次世界大战结束以后,一些国际组织(如世界银行)促使成本—收益分析应用到特殊项目活动、欠发达国家和工业化国家的国家级项目中(如果要了解这些年来效率分析在联邦政府中的应用,参见 Nelson, 1987)。

社会项目领域的成本—收益和成本—绩效分析与商界一样,成本总是要与

收益进行比较。例如,一家计算机公司在考虑让自己的个人电脑与主要竞争对手产品兼容的时候,最关心的就是成本与收益之间的关系。同样,小餐馆老板关心是,为了提高利润,是否应该为进餐者提供音乐或增加午餐特色菜。

用效率来判断社会干预的介入情况(商业术语是利润率)已经获得了广泛的认可。然而,"正确"计算社会项目成本—收益和成本—绩效的程序与方法却颇有争议(Eddy, 1992; Zerbe, 1998)。正如我们要讨论的一样,这样的争议与以下因素的多元作用有关:包括对分析程序的不熟悉,许多社会项目不愿意强调产出的货币价值,不愿意抛弃项目创建时所秉承的精神。因此,使用成本—收益或成本—绩效分析方法的评估者必须了解在特定领域使用效率分析会遇到的特殊问题,并应该了解成本—收益和成本—绩效分析的局限性(有关效率评估分析的综合讨论,参见 Gramblin, 1990; Nas, 1996; Yates, 1996)。

效率分析的重要概念

成本—收益和成本—绩效分析既是概念性的观点,也是复杂的技术程序。从概念性角度来看,或许效率分析最大的价值在于迫使我们对成本和收益状况进行考虑。如果是纯粹的社会项目,对实际或估计成本以及对已知或预期收益的确定和比较是没有意义的。大多数其他类型的评估关注的主要是收益问题。进而言之,效率分析在干预利用方面提供了可比较的视角。由于社会项目都在资源紧缺条件下运作,就不可避免地要对各种利用方式进行比较和判断。几乎无一例外的是,要维持和继续对项目的支持,就得满足政策决策者和投资者对项目判断的底线(例如,货币收益或其他等价物),以证明项目的正当性。

在这个领域中,非常有趣的一个决策例子是一家大银行为职工小孩提供日托的报告(参见专栏 11—A)。正如报告所描述的,尽管进行效率分析有一定的难度,尽管所获得的结果差强人意,这样的结果仍然能够提供证据,支持这类公司资助社会项目。专栏 11—A 的原始报告还讨论了各种商业渠道建立的保护健康项目、儿童日托中心、午间教育项目。在每个案例中,成本收益比较都是公司决策的基础。

专栏 11—A 银行提供儿童日托的成本节省

1987 年 1 月,联合银行在洛杉矶开了一家新的中心。然而,这一家中心并不向外借贷,也不处理有关资金的问题,而是照料儿童。

这家提供日托服务的机构位于银行货币园运作中心。1987 年,联合银行为这个中心总共提供了 105 000 美元。而这个项目为银行节省了 232 000 美元。当然,日托中心本身并没有什么特别之处,特别的只是 232 000 美元。这个数目是研究结果的一部分,该研究想告诉公司的是:公司的投资和政策(日托帮助、健康计划、哺乳离岗、弹性工作计划)正在获得货币收益。

联合银行的研究(试图涉及许多其他评估没有涉及的问题)对日托项目产生的效应做了进一步

的揭示。这个研究在中心开业的一年之前就已经展开,这样就使研究者能够进行更多的比较统计控制。在看了这个项目收益之后,联合银行同意为中心花费430 000美元。

利用银行人力资源部提供的数据,布鲁德(Sandra Burud,一位在加州帕萨德纳儿童服务中心的顾问)将项目实施第一年中出现的旷工、替岗、哺乳离岗时间与一年前的同样因素进行比较。她比较观察了87位使用中心设施的人员、105位与中心使用者类似的比较组人员,以及全体雇员。

她的结论是:按照最保守的估计,日托中心为银行一年节省的资金在138 000到232 000美元之间。布鲁德女士说,仅替岗而言,节省的资金就达到63 000~157 000美元,其依据是:使用日托中心的人员替岗率为2.2%,而控制组为9.5%;整个银行为18%。

她同时还计算避免旷工所节省的资金为35 000美元。日托中心的使用人员旷工时间比比较组平均少1.7天,哺乳离岗时间比其他雇员少1.2周。根据媒体对中心的报道,布鲁德女士认为,银行因此还获得了折合40 000美元的免费广告。

尽管测量的方式极其复杂,她认为研究结果有力地反驳了"把照料儿童简单化的观点","因为这不是那种仅仅自我感觉良好的项目,它不仅是一种员工福利也是一种管理工具"。

资料来源 J. Solomon, "Companies Try Measuring Cost Savings From New Types of Corporate Benefits," *Wall Street Journal*, December 29, 1988, p. B1. Reprinted by permission of *The Wall Street Journal*, Dow Jones & Company, Inc.

尽管很有价值,然而,在许多评估中,由于某些原因,正式的、完整的效率分析既不切实际也不明智。首先,所要求的技术程序或许在评估规划资源的能力范围之外,或者没有项目人员能使用这种复杂的方法,或者因为干预的结果极其微小或相当大而没有必要进行这类分析。第二,把某种投入和产出测量赋予经济价值会导致政治或道德争议,这样的争议会使另一些有用的评估结果潜在利用最小化。第三,用效率术语将评估研究的结果表述出来需要考虑主办方、项目各方、目标人群以及评估者本身所坚持的不同成本和收益观点(即**会计视角**,Accounting perspective)。对会计视角的依赖至少对某些项目方来说是难以理解的,这样,又一次将评估的适当性和有用性抹杀了(我们将在这一章详细地讨论会计视角)。

此外,效率分析的基本条件(所依赖的未经检测的假设或进行成本—收益或成本—绩效计算的数据)也许不能得到完全满足。甚至最倡导效率分析的人也承认,通常情况下,没有唯一"正确的"分析。在一些应用中,在运用分析性和概念性模型及其基本假设时,分析结果或许根本就对有价值的变异不敏感,进而使得结果不能令人接受。

尽管我们想强调在使用成本—收益和成本—绩效分析结果时一定要有所注意,有时甚至怀疑其可信度,但是,这种分析的确能为估计项目效率提供一种可重复的、理性的方法。不过,即使是效率分析的倡导者也很少认为效率分析是项目决策的唯一决定性因素。不管怎样,对决策而言,效率分析是很有价值的。

事前和事后效率分析

效率分析一般主要包括:①在规划设计阶段,根据预期成本和收益进行**事前**

效率分析(Ex ante efficiency analysis)。②在一个项目实施一段时间之后，对其进行回溯性的**事后效率评估**(Ex post efficiency analysis)，通过影响评估来衡量实施的有效性，以此决定项目是否进一步扩展或者保持。

在规划设计阶段，会根据预期成本和收益进行事前效率分析。当然，即使这样的分析仅仅只是一种推测，也必须要假设一个正面净影响值。同样，也要对项目提供和送达服务的成本进行估计。在某些情况下，如果已经有了探索性项目(或在其他地方有相同的项目)或项目的实施很简单，那么，对投入和影响的估计就会很确定。但是，由于全面或部分事前分析的基础并不是经验资料，所以对净收益的估计总会有所出入。事实上，在事前分析中，对投入和产出的估计历来都是有争议的议题。

对于一旦开始实施就很难放弃或在时间和经费方面需要得到共识的项目而言，事前的成本—收益分析尤其重要。例如，如果要在新泽西海岸建起新堤、为人们提供新的娱乐场所，那么只要建了海堤，就很难再放弃该项目；因此，有必要对这种方式和用其他方式提供娱乐场所的成本与收益进行比较，或者把提供娱乐场所和用同样资源建立其他社会项目的成本与收益进行比较。

因此，在许多情况下，如果项目主办方启动和维持一个项目需要充分的资源，那么，决策的产生就需要事前成本—收益分析。专栏 11—B 就说明了这样的情况，即对健康照料人员进行艾滋病毒(HIV)检测。据说，即使一次外科手术或换牙都有可能给病人传染 HIV/AIDS 病毒，后果非常严重且令人关注，但是，要在全国范围内对大量的健康照料人员进行 HIV 检测需要大量的花费。在项目实施之前，明智的做法是进行一些估计：相对于一定数量的人感染 HIV 而言，进行检测的费用大概如何。即使这种估计比较粗略，也会很有价值。专栏 11—B 的分析显示：在最有威胁的项目方案中，任何合理的政策选项都需要高额费用。而且，根据现有信息进行的估计也不确定。由于估计出来的费用高昂且不确定，决策者明智的做法就是——小心从事，直到获得更好的估计。

专栏 11—B 为健康照料人员安排 HIV 测试的事前成本—绩效分析

1994 年，菲利普和其他人一起研究了对健康照料人员(医生、外科大夫、牙医)进行 HIV 检测的成本—绩效。所考虑的政策选项有①强制性的和②自愿的检查；而对检测结果阳性的人，则要①与照料工作隔离，②约束其行为，③向病人告知(说明照料人员 HIV 呈阳性)。

在这个研究中，对成本的计算主要是通过文献研究和专家咨询。成本估计包括三个部分：①咨询和检测成本，②对早期 HIV 检测呈阳性的治疗成本，③每个感染人员的治疗成本。估计成本的方法就是将③从① + ②中分离出来。

根据高、中、低 HIV 感染的危险性，对所有可选方案进行分析，结论是：对健康照料人员进行一次性强制检测和对其中 HIV 阳性的人员强制限制其行为的方案，就成本—绩效而言，是最优的。尽管这已经是最低的成本，但对外科医生而言，每例感染的花费为 291 000 美元，对于牙医则为500 000 美元。考虑到高额成本以及强制限制病毒感染人员行为在政治上的困难，因此，这并不是一项可以考虑的政策选项。

分析还发现，成本—绩效估计对感染危险性的变化非常敏感，对医生(相对于牙医)的实践模式

也相当敏感。每例感染的累进成本也不相同,感染危险性较低的牙医为4.77亿美元,而感染危险性较高的外科医生则少81 000美元。

由于高额的成本和估计的不确定性,作者的结论如下:"基于伦理的、社会的和公众健康实践的考虑,如果对成本—绩效估计没有更大的确定性,就不应该执行强制检测政策。"

资料来源:Tevfik F. Nas, *Cost-Benefit Analysis: Theory and Application* (Thousand Oaks, CA: Sage, 1996), pp. 191-192. Original study was K. A. Phillips, R. A. Lowe, J. G. Kahn, P. Lurie, A. L. Avins, and D. Ciccarone, "The Cost Effectiveness of HIV Testing of Physicians and Dentists in the United States," *Journal of the American Medical Association*, 1994, 271: 851-858.

由于在社会项目中对事前分析使用的并不充分,许多社会项目在启动和修改时,都没有关注成本—收益和成本—绩效分析。例如,如果防止蛀牙项目每年每个儿童的成本为200美元,而项目预计每年将使每个儿童减少半颗蛀牙,即使这种方式能奏效,这样的项目也很难被接受。毕竟,这个成本是牙医填一颗蛀牙费用的4倍左右。因此,效率分析能轻易地帮助决策者确定是否实施项目。

在更多情况下,效率分析往往发生在影响评估之后,因为那时已经知道了项目的净效果。这种事后成本—收益和成本—绩效评估所要获得的就是项目的绝对或相对效率,或者两者都有。在所有情况下,分析考虑的就是与净效果比较,看干预所花费的成本是否值得。

如果考虑的是绝对效率,需要判断的就是,相对于项目的效果(收益或结果)而言,项目成本是否值得。例如,如果成本—收益分析的结果显示,在综合商场中每花1美元用于防止偷窃就能减少2美元的商品被盗损失,那么,这个项目在经济上就是值得的。但是,作为一种分析方法,成本—绩效研究也可能会向我们展示:这个项目每减少一起偷窃就需要花上50美元的成本。

如果考虑的是相对效率,需要判断的是,一个项目相对于另一个项目的收益差异。例如,通过一个电脑教育项目来提高学童的阅读成绩,使其达到一级阅读水平所花费的成本,如果与通过同龄人辅导项目达到同等水平所花费的成本进行比较,就可以看出哪种项目方式能更加有效。在事后效率分析中,对成本和收益评价的基础就是前些章节中已经谈及的项目督导和影响评估类型的研究。

成本—收益和成本—绩效分析

显然,除了经济效率以外的许多考虑都可以用于政策决策、计划和项目执行,但是,由于资源总是稀缺的,所以经济效率总是十分重要。成本—收益和成本—绩效分析可以激励评估者去了解项目成本;令人惊讶的是,许多评估者对成本概念很少注意,并对项目所需的资源和成本的复杂性知之甚少。与此相应的是,项目成本对于项目各方是否接受或修改项目,却十分重要;因此,对成本的关注会使评估者获得项目各方的更多合作和支持。

成本—收益分析需要对一个项目的收益进行估计,收益包括有形的和无形的两种形式;并且还要对项目的成本进行估计,包括直接的和间接的两类成本。一旦涉及具体情况,就要把收益和成本转换成一般测量,通常用货币形式来

表示。

成本—收益分析需要采用特殊的经济视角,此外,还需要一些假设,据此才能把投入和产出转换成货币形式。正如我们已经注意到的,对于投入和产出"正确的"货币转换程序,现在还有许多争论。很显然,成本和收益测量定义所隐含的假设,对分析结论有很大影响。因此,分析者至少应该说明用于分析的假设基础。

在通常情况下,分析者所做的比这要多。根据不同的假设,他们或许对同一个项目进行各种不同角度的敏感性分析(改变重要假设),从而检验研究结果对于不同分析假设的敏感性。敏感性分析(能够提出重要的假设,并能估计项目的产出)是好的效率研究的核心特征。事实上,基于丰富的成本和收益信息的、正式的效率研究的重要优势在于,假设和程序是公开的、可以检验的。例如,使用事前成本—收益分析方法,如果对一个意图减少累计犯罪的项目成本节约进行分析,就必须能估计到改造被捕者的司法程序成本。项目所节约的成本依赖于成本值设定本身,成本是被假定为 5 000,10 000 美元或者其他数字,对于计算结果影响很大。

总的来说,对产出的货币转换要比对投入的转换有争议得多。在技术性和工业项目领域,应用成本—收益分析技术的争议较少,因为比较容易将收益和成本转换为货币形式。譬如用于降低用电成本的工程项目、便于货物运输的高速公路建设项目、用于提高农作物产量的灌溉项目。在社会项目中,要用货币方式将收益估计出来,常常要困难得多,因为在这些项目中,只有一部分项目的投入和产出可以进行货币转化。例如,对教育项目而言,可以把未来的职业收入用货币形式表达而不至于引起大的争论。但在生育控制或健康服务这样的社会项目中,情况就要复杂得多,因为必须把人们的生活赋予货币价值,才能对项目进行货币形式的表达(Jones-Lee, 1994; Mishan, 1988)。

所以,基本的原则就是——成本—收益分析试图依据边际社会价值,对投入和产出量进行价值转换。在很多情况下,类似于提供某种医疗服务所需的成本、为减少汽车耗油量而提供新引擎服务所得到的货币收益以及市场价格,都能进行价值转换。另一方面,当物品或服务没有成为交易对象时,进行价值转换的困难就大多了。

由于对产出定价本身的争议性,在很多情况下,尤其在考虑服务时,成本—绩效分析往往比成本—收益分析更合适。成本—绩效分析只需要将项目的成本转化为货币形式;而收益则用产出本身的单元来表示。例如,为乡村小学儿童提供免费书本,项目的成本—绩效分析表达形式为,每 1 000 美元项目成本能够提高几分儿童阅读平均成绩。

这样,对于成本—绩效分析而言,效率被表达为给定结果的成本。这就是说,对项目的效率分析可以用这样的方式来表达,即每单位产出所消耗的资源或成本。这种分析方法在比较不同项目效率的研究中,显得尤为有用。例如,就教育项目而言,就是具体教育项目的每一考试分数所花费的成本。专栏 11—C 显

示了小学教育中儿童数学与阅读能力提高的成本。分析表明,按照投入100美元的结果,一对一的人工教育所产生影响要高于其他的方法。令人惊讶的是,在成本—绩效方面,同伴教育的影响甚至比计算机辅助程序的效果要好。

专栏11—6 计算机辅助教育的成本—绩效分析

为帮助决策者考虑各种方法对提高小学生数学和阅读水平的影响,研究者对计算机辅助教育(CAI)进行了成本—绩效分析,并与其他三种方法进行了比较。研究结论与某些已有的看法正好相反。按照成本—绩效标准,尽管CAI出色地完成了任务,但效果还是不如一对一的人工教育。令人惊讶的是,传统的、劳动密集型的方法(人工教育)比电子干预的效果(成本—绩效)要好得多。从成本—绩效的角度来看,尽管有许多呼声要求进行教育改革,但增加课内教学时间的效率说明,至少对帮助提高阅读和数学能力方面而言,增加课堂时间不是一个好的选择(见表格)。

为了估计不同方法的成本—绩效,研究者们确定了各种方法在提高数学和阅读考试成绩方面的标准差。接着,确定了每种方法的成本,计算每个孩子每花100美元所取得的成绩。数学和阅读成绩提高的结果如下表。

针对两门课进行四种干预的成本一绩效比率的平均值(在每门课中为每个孩子每花100美元,数学和阅读能力成绩提高的平均值)

改革方式	成本—绩效比率
年龄交互型人工教育	
同龄人和成年人结合项目	.22
同龄人部分	.34
成年人部分	.07
电脑辅助教育	.15
班级缩小的规模	
从35到30	.11
从30到25	.09
从25到20	.08
从35到20	.09
教育时间的增加	.09

资料来源:H. M. Levin, G. V. Glass, and G. R. Meister, "Cost-Effectiveness of Computer-Assisted Instruction," *Evaluation Reviews*, 1987, 11(1): 50-72.

事前成本—绩效分析能根据一定成本下的效果值对项目进行排队和比较。在事后成本—绩效分析中,用于比较的就是实际的成本和影响(包括投入和产出),并可以和原来的估计及假设进行比较。此外,回溯性分析能提供有用的理解、经验甚至方法及程序,以用于未来的项目研究。然而,对产出和成本之间的比较需要项目有同类的产出。如果项目的产出不同,譬如帮助病人尽快康复的医疗项目和帮助

提高阅读水平的项目,要将两者的产出进行比较是很困难的。如果要比较,那么每个病人减少两天卧床时间与阅读水平提高之间比较的“价值”在哪里?

效率分析的使用

效率分析,至少是事后分析,是影响评估的延伸,而不是影响评估的一种。因为,对货币收益或具体效果估计,都要涉及项目的净影响。如果项目的影响不可知或无法估计,那么要进行成本—收益或成本—绩效计算就不可能。对无效的项目(也就是说,如果影响评估发现没有显著的净效果)进行这样的分析是没有意义的。如果没有对项目效果进行合理的估计,对于正在实施的或已经完成的项目进行效率分析,同样也是愚蠢的。当效率分析被应用到有效的项目时,对于决策者考虑支持某个项目而不是另一个项目、从绝对意义上考虑某个项目的成本与产出相比较是否值得推进、考察某个项目在不同时间地点的利用等内容而言,就非常有用。此外,效率分析在决定项目收益所达到的水平方面起着十分重要的作用,并能为改善项目提供有效信息(Yates, 1996)。

成本—收益分析

有了关于效率分析的最基本概念,现在就可以转而讨论效率分析的具体实施。因为很多基本程序都是类似的,所以,我们将只详细讨论成本—收益分析,而比较简单地阐述成本—绩效分析方法,两部分内容可以相互参照。首先,我们必须对这两种分析方法的一个共同主题——成本数据的搜集,进行必要的说明。

汇集成本数据

显然,成本数据对于项目效率的测量十分重要。在事前分析中,必须根据以往类似项目的成本或项目过程成本的知识,对项目实施所需成本进行初步的估计。对事后分析来讲,则很有必要核查、分析项目的财政预算,分清每个项目阶段的财政支出和花费在项目对象及其他机构上的成本。

成本数据的有效来源包括下面几个方面:

- 机构财务记录:涉及项目工作人员工资、办公空间租用费用、客户补贴、物资供应、维修费用、商业服务开销等内容。
- 目标对象成本估计:包括项目活动中客户所耗费的时间成本及交通成本等(很明显,必须估计这些成本)。
- 协作机构:如果一个项目中有诸如学校、健康诊所或者政府机构之类的合作者,就要从这些机构获得成本数据。

必须指出的是,财务记录并不总是容易弄清楚,评估者可能必须向专业会计人员寻求帮助。

面对一个社会项目,列出其成本数据的清单通常是很有用的。专栏 11—D

是一个工作表,列举了一个具体干预项目的各种成本。这个项目研究的是“与科学家的接触对高中生科学职业追求兴趣的影响”。表中列出了与项目相关的几个成本衍生要素:

专栏 11—D　一个假定项目的年均成本核算清单

“星期六科学家”是高中生跟当地大学教授及高中自然科学课程老师进行互动的一个项目。活动在每月内的两个周六进行,目的在于激发学生投身科学事业的热情和兴趣,并将他们带到科学研究的前沿。下面这个表展示了不同层面的政府、学校和参与项目的学生及其父母所分别承担的各种项目成本。

成本类型	总成本	地方学校成本	州政府成本	大学成本	学生及家长成本
1. 人力资源					
两个高中老师	9 000	9 000			
两个大学教授	14 400			14 400	
两个父母作为助手(志愿者)	3 600			3 600	
2. 设施					
高中实验室和教室	2 000	2 000			
3. 物品和设备					
照相器材	400	400			
科学实验材料	500	250		250	
实验室设备	500			500	
4. 其他					
维护和看门服务	1 500	1 500			
保险	1 800	1 800			
其他零散费用	900	900			
5. 委托人投入					
交通费用:时间、交通工具成本	625				625
6. 各要素总成本	35 255	15 850	0	15 150	4 225
7. 使用者的费用		-1 000			1 000
8. 其他现金补助		-7 500	7 500		
9. 净成本	35 225	7 350	7 500	15 150	5 225

资料来源:Henry M. Levin and Patrick J. McEwan, *Cost-Effectiveness Analysis*, 2nd ed., Table 5.2. Thousand Oaks, CA: Sage, 2001.

会计视角

为了进行成本—收益分析，首先必须确定在进行成本与收益计算时应该考虑的问题。将收益和成本转换成具体的、可测量的货币形式的基础是什么？简而言之，谁将承担成本、获取收益？对收益和成本下定义必须从某个视角出发，因为视角的混乱会造成具体分析的含糊性、重复或重叠。当然，对一个项目可以进行多种成本—收益分析，每种分析通过不同的角度来完成。基于不同观点的分析能为成本和收益之间的比较提供有价值的信息。前面我们已经指出，进行成本和收益估计时，需要确定一个会计视角。总体上看来，对于社会项目而言，应该考虑三种不同的会计视角：①单个参与者或对象；②项目主办方；③项目中的公众集合体，或者是社会环境。

单个对象会计视角基于项目干预对象来考虑问题，分析单元可以是个体、组织或群体。使用单个对象会计视角的成本—收益分析所获得的收益—成本比值（净收益）比其他方式获得的要高。换句话说，如果主办方或者社会承担了一项成功干预项目的成本，那么单个项目参与者就是最大的受益者。例如，教育项目让参与者承担的成本就很少。事实上，参与者要承担的成本就是他们参加项目所花费的时间，因为书本和资料已经为他们准备好了。而且，如果安排的时间是晚上和下午，那么参与者就不会有收入方面的损失。同时，参与者获得的收益或许包括收入的改善、教育水平的提高、更高的工作满意度、更好的职业选择以及通过项目获得的转移收入。

项目主办方会计视角用筹资的观点来看待收益的价值和成本因素。经费来源可以是私人机构、政府部门、营利性公司。从这个观点看来，成本—收益分析在很大程度上类似于通常提到的私营利润率分析（private profitability analysis）。这就是说，从这个观点出发的分析就是要揭示主办方为项目所付出的成本及想获得的收益（或“利润”）。

主办方会计视角最适合这样的情境：如果主办方面对的是硬预算（也就是说，不可能再找到其他经费来源），而又必须在可选项目中进行选择。在这种情况下，如果项目主办方是县政府，也许会更赞成包括学生补贴的假期教育项目，因为这种项目会减少公共辅助成本和类似津贴（因为参加假期教育项目的人或许已经收到了来自收入维持基金组织的支持）。同样，如果参与者将来的收入由于受到了教育而有所提高的话，他们直接和间接要交的税金也会相应地提高，这也是从项目主办方角度去计算的收益。政府主办方花费的成本包括项目运作成本、管理成本、辅导成本、供给成本、设施成本以及在培训期间为参与者提供的津贴。专栏 11—E 显示了精神健康系统的成本—收益计算，即通过为患有精神疾病同时又有物质滥用的人提供专门服务所节约的费用。

专栏11—1 提供专门化双重诊治精神健康项目的成本与节约分析

在一般的精神健康或者物质滥用服务中,对患有精神疾病同时又有物质滥用的人(双重诊断)的治疗,不仅非常困难,而且成本高昂。为他们提供特殊的双重诊治方案会使结果有所改善,但服务经费也会相应提高。然而,双重诊治产出的改善能减少对精神健康服务的需求量,相应地也会减少项目成本。因此,从精神健康系统决策者的角度来看,一个重要的问题就是,在这个特殊项目上花费的成本是否能从对服务的需求减少这个环节上得到补偿。

为了得到答案,评估研究者随机抽查了132名病人,将他们分配到三个具体双重诊治项目中,并对产出和成本进行评价。“控制”项目的基础是12步骤康复模式,进行的是精神健康系统的“通常治疗”。另外,也请酒精和麻醉品管制委员会为项目实施提供支持与服务,帮助成瘾患者康复。第二个更为加强的行为技能模式是认知—行为治疗,关注社会和独立生活技能、防止复吸。第三个不那么加强护理的模式就是案例管理,允许医生在日常生活或法律领域为病人提供个人的帮助,因此,所能管理的案例数量也会减少。

按照病人康复表现和症状来看,行为技能模式获得了最大的正面效果,但同时,成本也最高。为了进一步考察成本情况,评估者分四个阶段考察了三个项目的服务利用和服务成本,即双重诊治之前6个月、之后6个月、之后12个月和之后18个月。

精神治疗服务成本被分成两类:支持性服务和加强性服务。支持性服务包括案例管理、门诊就医、药物治疗、日常服务以及其他为病人安排的常规性服务。加强性服务包括为重患者准备成本更高的治疗服务,例如,住院病人服务、职业护士照料、住地治疗以及突发情况下的诊治。

由于要提供额外的资源,所以双重诊治的支持性服务成本应该会上升。同样,由于加强性服务的减少,精神健康系统的成本应该有大量的节省。这样,成本分析关注的就是支持性服务成本增加与加强性服务成本减少之间的比较。表格显示的结果是,在项目开始后的第6个月和第18个月之间服务利用成本的变化情况。

按照最初预期,除了案例管理项目以外,支持性服务成本在项目实施之后应大幅增加,然而实际上,支持服务总成本(与基本服务比较)却有所下降。另一方面,支持性服务成本上升最快的就是行为技能项目。

同样,与所有专门项目的基本服务比较,加强性服务的成本也下降了。然而,行为技能项目对病人行为和症状更大的影响并没有减少服务利用和相应的成本。事实上,是12步骤的普通护理为随后的加强性服务最大限度地降低了成本。然而,案例管理项目没有造成成本的大幅下降,因为与12步骤项目比较案例管理的成本原本就比较低。额外的分析表明,这些项目同时也节约了医疗成本、司法成本以及参与者家庭的成本。

因此,仅仅就精神健康系统的成本节省而言,12个步骤和案例管理项目所节省的比所耗费的开支要多。事实上,根据成本分析估计,在实施项目18个月后,每1美元的项目投入,就能得到9美元的节省。而且,对案例管理项目的实施能直接使支持性服务成本下降,即不需要额外的投资了。然而,另一方面,行为技能项目使精神健康系统直接受损。每投入1美元,只能得到0.53美元的节省。

在项目执行18个月后服务每个病人所需成本的变化情况　　单位:美元

	12步骤项目	行为技能	案例管理
精神健康治疗支持性成本的变化(a)	+728	+1 146	-370
精神健康治疗加强性成本的变化(b)	-6 589	-612	-3 291
(b)与(a)之间的比值	9.05	0.53	8.89

资料来源:Jeanette M. Jerell and The-Wei Hu, "Estimating the Cost Impact of Three Dual Diagnosis Treatment Programs," *Evaluation Review*, 1996, 20(2): 160-180.

集体会计视角用集体或整个社会的观点来衡量项目,通常用总收益来表达。这是最综合的视角,同时也是最复杂的、最难应用的视角。从整体社会的角度来看,意味着对项目而言,具体努力产生的都是**次生效果**(Secondary effects)(不直接牵扯干预对象本身,体现在其他一些群体身上的效果,包括有益的或者有害的两个方向)。例如,一个教育培训项目可能会具有溢出效应,会波及对亲属、邻居、参与者的朋友的教育。而更为经常提及的例子,则是工业的和技术的项目所带来的负面影响,包括污染、噪音、交通问题和对动植物的危害。而且,根据现有文献,集体的成本—收益分析范围已经扩大,包括对平等的考虑,也就是说,在不同下属部门的**配置性效果**(Distributional effects)。例如,从集体的角度来看,曾经失业6个月或更长时间的少数族裔每挣1美元就会被视为具有"双重收益(double benefit)"并计入分析。

专栏11—F显示的是,从集体视角出发需要考虑的收益。在专栏中,格雷及其同事(Gray, 1991)试图对各种刑罚效率的成本—收益分析进行整合。正如专栏11—F显示的,收益有各种形式。也正如文章所要谨慎说明的,尽管估计的精度有一些局限,但研究结果对于法官和其他司法专家为各种刑罚进行判断提供了重要信息,对于他们了解不同类型审判的成本很有价值。

专栏11—F　适当刑罚方法的成本与收益

使用适当刑罚方法能起到控制犯罪率的作用,对确认有罪的罪犯,其判决的执行(通常有三种方法:拘留所、牢房和缓刑)不仅要考虑成本,而且要考虑收益。每种刑罚对社会都有不同的收益。主要的有取消资格,即把罪犯从社会集体中分离出来送到牢房或是拘留所;通过对犯罪行为的可见吓阻来阻止潜在的犯罪;还有康复,通过再社会化进行行为矫正。因为拘留处罚维持的时间通常很短,这样,与监狱处罚相比,"取消资格"产生的价值稍微小些,尽管如此,因为不会有人喜欢呆在拘留所里,所以"吓阻"带来的社会效益是监狱的5~6倍。

格雷及其同事试图估计每种刑罚所带来的社会收益的货币价值(见表格)。然而,从平均值来看,缓刑带来的净收益要比拘留所的大,而后者又比监狱带来的收益大,因犯罪类型和环境不同,每种收益的权重也不一样。例如,对入室行窃处罚的成本(包括受害人的损失、警察局调查成本、逮捕罪犯所花费用以及法庭审判成本)总共将近5 000美元,这样,利用长时间的监狱处罚来惩罚惯犯就能够使"取消资格"的收益最大化。与此相反的是,要抓获销赃罪犯的成本是2 000美元,那么,短时

间的拘留甚至缓刑也能达到最好的效果。

每个罪犯每年花费的社会矫正成本和产生的收益估计 单位:美元/处罚(平均值)

	取消资格收益	康复收益	吓阻收益	成本	纯利益
监押	+6 732	10 356	+6 113	-10 435	-7 946
拘留	+774	-5 410	+5 094	-2 772	-2 315
缓刑	0	-2 874	+5 725	-1 675	+1 176

资料来源:T. Gray, C. R. Larsen, P. Haynes, and K. W. Olson, "Using Cost-Benefit Analysis to Evaluate Correctional Sentences," *Evaluation Review*, 1991, 15(4): 471-481.

从集体角度出发,成本—收益分析的组成部分包括单个对象、项目主办方计算成本和收益构成的大部分,但对每项内容的估价不同。例如,一个项目的集体成本包括由社区提供基金的**机会成本**(Opportunity costs)。这样的机会成本明显不同于作为项目参与者的机会成本。集体成本同时还包括设施、设备、人力资源,等等,但从项目主办方的角度来看,这些成本的价值就不同了。最后,这些成本还不包括转移支付,因为这是集体的收益,而两者不能共存,或计为成本或计为收益。

显然,选择怎样的会计视角取决于作为评估结果关注者的项目各方或项目主办方。因此,对会计视角的选择是一种政治决策。例如,如果是只对医院护理成本感兴趣的私人机构聘请的分析者,当然会采用项目主办方视角。如果从单个对象视角出发,分析者就会忽视甚至根本不理会对项目主办方而言能够获得最大净收益的成本控制(不等于成本节约),并将其计为负面收益。因为高昂的护理费用,家庭成员不得不离岗从而为病人提供专门的服务,因为医院要求总有人陪伴在病人左右。

总的来说,集体会计视角在政治方面是最中性化的。如果分析者能合理使用这种方法,那么,从单个对象或项目主办方得到的信息,就应该作为成本和收益被囊括进来。另一种方法是从多个角度进行成本—收益分析。问题是,和其他评估活动一样,成本—收益分析也有政治特征。

专栏 11—G 显示了从不同会计视角出发进行成本—收益分析所包含的基本组成部分(项目主办方在这种情况下是一个政府部门)。当然,这里并没有囊括全部内容,仅仅只是一个参考。在实际分析中,具体内容总会有所不同。

专栏 11—G 从不同角度进行成本—收益分析的要素（以培训项目为例）

	单个对象（目标人群）	项目主办方（政府）	集体（总体上的组织）
收益	纯收入的增长（税后）	税金的增长	总收入的增长（税前）
	额外收益（如直接转移、附加福利和非经济收益）	公共辅助和其他津贴支出的减少	其他收入的增加（如附加福利，不包括直接转移）
		项目已完成工作的价值（以市场价格计算的工资和福利）	不再适用的替代性项目支出减少
			项目已完成工作的价值（以市场价格计算的工资和福利）
成本	机会成本（流失的纯收入）	失去的税金	机会成本（流失的总收入）
	不再利用的直接津贴损失（备选社会项目）	项目成本（如资金、管理、辅导、直接津贴	项目成本（不包括直接津贴或转移支付）
	与参与者相关的成本（如酬金、资料）		

资料来源：Jeanette M. Jerell and Teh-Wei Hu, "Estimating the Cost Impact of Three Dual Diagnosis Treatment Programs," *Evaluation Review*, 1996, 20(2): 160-180.

专栏 11—H 提供了一个简单的假设性例子，从三种不同的会计视角计算培训项目的成本—收益。在这里，使用货币数字主要是为了简化分析过程，真正的分析要对相关测量问题做更复杂的处理。请注意，在计算过程中，某些内容可以在一种计算方式中被划入成本部分而在另一种计算方式中则被列入收益范畴。对收益和成本的差值（即净收益），不同的视角也有不同的算法。

在某些情况下，要进行多种分析。例如，如果政府部门和私有基金共同作为项目主办方，就要对两者分别分析，从而判断出各自的投入回报。同时，分析者也许想计算不同对象的成本和收益，譬如项目直接的和间接的目标人群。例如，许多社区都会为驻地工业企业提供税收优惠，目的就是为本地居民提供就业机会。成本与收益的比较可以从雇主、雇员角度去算，也可以从社区每个居民的角度去算，他们薪水的增加会导致所交税金的上升。当然，也可以包括其他类型的分解。例如，从集体视角出发就不计算直接津贴（既是成本，又是收益），因为两者之间始终是平衡的；然而，在某种条件下，津贴实际经济收益会小于成本。

如果想要更为细致地了解效率分析的实际操作方法，可以参考格林伯格和阿彭泽勒（Greenberg and Appenzeller, 1998）的著作，这本书以每年的雇工培训项目和工作福利项目为例，详细地介绍了效率分析方法的每一步操作过程。

专栏 11—11 从不同的会计角度进行成本收益计算(假设的案例)

单价:$元

收益/成本	
(1)被培训人员收入的改善(税前)	100 000
(2)被培训人员收入的改善(税后)	80 000
(3)在培训期间已完成工作的价值	10 000
(4)项目花费的设施和人力资源成本	50 000
(5)项目花费的设备和供给成本	5 000
(6)被培训人员的津贴(直接转移支付)	12 000
(7)被培训人员放弃的收入(税前)	11 000
(8)受培训人员放弃的收入(税后)	9 000
(9)损失的税金(7)—(8)	2 000

	单个对象	项目主办方	集体
收益	(2) 80 000	(1)—(2) 20 000	(1) 100 000
	(6) 12 000	(3) 10 000	(3) 10 000
	92 000	30 000	110 000
成本	(8) 9 000	(4) 50 000	(4) 50 000
		(5) 5 000	(5) 5 000
		(6) 12 000	(7) 11 000
		2 000	66 000
		69 000	
净收益[a]	83 000	-39 000	44 000

a. 请注意:净社会收益可以分成两部分,受训人员的净收益加上政府的净收益,在这种情况下,后者为负值 44 000 = 83 000 + (-39 000)

测量成本和收益

成本和收益的具体化、测量和估计(成本—收益分析的核心步骤)提出了两个截然不同的问题:第一,识别和测量项目的所有成本与收益,第二,用一般方式对所有成本和收益进行表达,也就是将其转换为货币形式。在事先评估中,成本和收益的识别与测量问题比较尖锐,因为那时只有对成本和影响的估计。然而,在事后成本—收益分析中,也有资料短缺的问题。对于许多社会项目而言,从评估获得的信息(即使是一系列的评估)也不足以满足回溯性成本—收益分析的需要。这样,评估往往仅能提供一些必要的信息,因此分析者必须经常使用另外的

资源或判断依据。

在许多社会项目中出现的第二个问题是，很难将成本和收益转换为货币形式。社会项目的产出常常不能用市场价格进行精确计算。例如，许多人会对生育控制项目的收益产生异议；或者一个通过培训来改善健康实践的项目，会因为项目各方的不同观点而无法转换为可以接受的货币形式。对文盲的价值怎么计算？在这种情况下，成本—绩效分析就会是一种合理的方法，因为这种分析不需要将获益用货币形式进行表达，只要有产出测量就可以了。

货币收益

因为把收益表达为货币形式有很多优点，所以人们总结了一些方法，用来将产出和收益用货币形式表达出来(Thompson, 1980)。以下是五种常用的方法：

1. **资金测量**。最不容易产生分歧的方法是估计直接货币收益。例如，如果在晚上将健康中心的开放时间延长两个小时能帮助减少目标人群每年旷工的时间(导致薪水减少几率的减小)10小时，那么，从个体会计角度看，年收益的计算方法是将平均工资乘以10个小时。

2. **市场评估**。另一种相对不会产生异议的方法是，将收益或项目影响按照市场价格转换为货币形式。如果社区的犯罪率减少50%，用低犯罪率社区的住房价格就可以计算出所获得的收益。

3. **计量经济评估**。较为复杂的方法就是对假定的产出或影响进行市场估价。例如，由于对犯罪行为恐惧的降低，导致企业效益增加，从而导致税收增加。这种结果多大程度上归因于项目干预的计算方式就是用低犯罪率社区的相关税收额进行税收增加估计。这种估计需要较为复杂的分析能力，分析者应该是接受过高级训练的经济分析师。

计量经济评估，如果运用当前多变量分析技术，就是一种较受欢迎的选择，因为多变量分析能够考虑变量的其他影响(在先前的例子中，由于对犯罪的害怕而导致税金的损失)。要做出高质量的计量经济模型需要很高的分析能力和假设能力。然而，计量经济分析，像其他所有方法一样，需要清楚地描述假设，以确保人们能够考察分析的基础。

4. **假设性问题**。这是一个稍有问题的方法，通过直接对对象提问获得对一些非货币收益的估价。例如，防止蛀牙项目的效果是平均40岁的人减少了一颗蛀牙；在调查的时候，人们会问，保住一颗整牙和填一颗蛀牙到底值多少钱。这样的估计就是假设人们的说法就代表了整牙的实际价值。很显然，这类假设价值是可以质疑的。

5. **观测政治选择**。最不稳定的方法就是从政治意义上进行价值估计。如果政府决策部门愿意欣赏这样的说法，资助高危婴儿医疗的项目节省了大笔医疗费用，每个婴儿大约节省40 000美元，那么，这个数字就能被用来衡量这个项目的货币价值。但是，如果政治选择非常复杂、琢磨不定，而且很不连贯，那么，这种方法就很不可靠。

总而言之,如果成本—收益分析要求有较强的可信性并能完整地反映经济效果的话,就必须包括所有相关部分。如果重要收益因为无法进行货币转换而被忽视,那么项目的效果看起来就不一定如预想的那样高;相反,如果忽略了某些成本,项目就会看起来更有效。不论对成本和收益估计太保守或太过分,都会形成对产出测量的误导。作为处理问题的方式之一,分析者往往会尽力把所有能够被转换的事项转换成货币形式,并将无法转换的事项列举出来。为了让项目能"进行下去",他们也将估计非货币收益的价值。

影子价格

成本和收益需要用不同的方式进行定义和估计,其基础是会计视角。然而,对于许多项目而言,产出并没有市场价格(例如,污染量的减少或者是家务活的减少),但是还是应该估计其价值。较好的方法就是**影子价格**,同样也被称为会计价格,为的是反应真正的市场价格,从而得到对成本和收益真正的估计。影子价格反应的是商品和服务的价格,为的是反应真正的收益和成本。有时,即使有实际价格,使用影子价格可能更为现实。例如,一个实验性项目的实施需要一个对建筑业了如指掌的经理。对于一个项目点而言,主办方或许能幸运地找到一位退休者,后者或许对该项目很感兴趣,也愿意参与,假设他的报酬是每年30 000美元。如果通过影响评估说明项目是成功的,那么在进行成本—收益分析时,最好使用影子价格,就是说,经理的薪水是50 000 美元,因为或许除了他以外,没有人对这样低的薪水感兴趣(Nas, 1996)。

机会成本

机会成本概念反应的事实是——在通常情况下,资源是有限的。个人和组织常常会从已有的资源配置方案中进行选择,这些选择同时也对决策者的活动和目标产生影响。每一种被选择的成本都能用被放弃的其他选择进行逆向测量。

尽管这个概念相对简单,但机会成本的实际估计却常常十分复杂。例如,警察部门决定为警察官员们付学费,让他们去学习心理学或社会工作,这样能帮助他们提高工作绩效。为了帮助这个项目获得资金,这个部门会使他们的警车在两个月之内进行额外的工作。在这种情况下,就可以用警察局额外的交通费用来测量机会成本。因为在很多情况下,只能通过对备用投资选项的假设结果进行机会成本估计。在效率分析中,这是一个有争议的领域。

次生效果(外部性)

项目还会产生外在的(附带的、意料之外的)的影响,这包括正面收益或有害影响。因为这样的效果并不在预期之中,如果不使用特殊的办法把这些包括进来,就会在成本—收益计算中被忽略掉,这是不合适的。例如,培训项目的次要效果或许是参与人的亲戚、邻居和朋友都受到了相应的教育。负面影响的常见

例子有，在工业和技术发展的同时，产生了诸如污染、噪音和交通问题，严重地影响了生物的生存。

对于许多社会项目而言，一般会产生两种次要产出：置换和真空效果。例如，教育或培训项目会使一批刚刚受过培训的人加入劳动力市场，与在职人员形成竞争，并可能取代他们（如，迫使他们离岗）。项目的参与者也许刚刚从某个职位上下来，也就是说，腾出了位置。

识别和测量次生效果或外部性是困难的。然而，一旦发现了，就应该将其计入成本—收益分析。

分布问题

从传统角度看，对社会项目绩效的判断所秉承的宗旨是，至少使一个人受益而没有人受害。从经济学角度来看，这就是帕累托标准（pareto criterion）。然而，成本—收益分析并不使用帕累托标准，而使用潜在帕累托标准。在这种标准下，获得的成果不仅有潜力弥补损失，而且还有剩余。这就是说，如果要估计（不一定进行检验）项目影响，应该考虑的是受益对象多于受害对象，用更为精确的方式表达就是，用总收益减去总损失之后获得的是正值。然而，在社会项目中，这种标准很难得到认同，尤其是那些依赖收入转移的项目。例如，降低十多岁年轻人的最低收入标准会增加他们的就业机会，同时也会减少年长人口的就业机会。

在通常情况下，问题不单单是输赢之间的关系，而是要在目标人群中保持服务公平性。如果要提高一个群体或社区的平均生活质量，这种情况就尤其突出。在成本—收益分析中，考虑平等和分布问题的基本方法是：如果项目起到了预期的积极作用，就通过系统加权的方式来进行调整。这样，如果降低年轻人最低收入标准的项目导致家庭收入下降，进而对家庭产生不利影响，就可以根据对家庭不利影响的程度来进行收入和损失加权。对社区而言，无论从平等角度还是从改善人们生活的角度，如果一些项目较其他的项目更为有价值，那么就应该给予这样的项目以更高的权重。

如何使用权重可以由相应的决策者来确定，其中显然会包括一些价值判断。权重的使用也可以根据经济原理和假设来确定。必须明确的是，在任何情况下，权重的使用都不可能是无偏的。当分析者要继续处理分布时，无疑要做进一步细致的分析。进行公平的成本—收益分析的一个折衷解决方案，就是首先检验项目的成本和收益是否满足潜在帕累托标准。如果满足，就可以进行分组计算。分组的标准可以是收入、学生学习成绩等。在分析类似于学校效果之类的问题时，分布问题尤其重要，因为其中的部分成本是纳税人支付的，但他们并没有获得直接收益。公立教育的收益主要落到了有孩子上学的家庭和不那么富裕的家庭，相应地，这些家庭所纳的税也较少。

折　扣

项目效率分析方法中另一个重要的、与成本和收益估计有关的因素是时间。

项目随时间演进而发展变化,成功的项目会在未来形成收益,有时甚至是在项目完成之后很长时间。有些项目收益还会伴随参与者终身。那么,评估者就要推断未来的影响、估计未来的收益,尤其是在项目收益与参与者收入挂钩的情况下。为了完成分析,事前评估常常需要对未来的状况进行推断。否则,评估的基础就会局限在有限的、可以得到项目绩效数据的时期。

因此,不同时点的成本和收益必须要用一般性方式进行表达,并可被普遍接受。换而言之,必须要考虑项目成本和收益的时间模式。可用的技能就是**折扣**,把随时间而减少的成本和收益用一般的货币形式表达,或将其换算成现价。例如,在项目开始的时候,成本往往是最高的,那时,许多资源都必须被调动起来;在项目结束时,这些资源要么渐渐变少,要么已被用尽。甚至在成本固定或收益是常数的情况下,不同时点的成本和收益也是不等价的。我们不能这样问,"在将来,我的投资会带来多少收益?"在经济实践中,通常的问法是,"将来产生的收益与现在的相比,其减少的程度如何?"对成本而言,道理也是一样。回答这些问题的基础是对利率或折扣率的假设,以及对时间的选择。专栏 11—I 为折扣提供了一个例子。

专栏 11—I 根据现价进行成本和收益折扣计算

折扣建立在一个简单的观念之上——现有资金而不是未来的资金量是估计的基础。在其他方面一样的前提下,现有资金可以存进银行去获得利息,也可以用来做其他投资。这样,现有资金在将来就会比其自身的面值高。换句话说,将来某个可支付的数额不如现有同样的数额资金值钱。

在概念上,折扣与利息正好相反,告诉我们现在要把多少资金放在一边,才能在将来获得某个固定数额。从数学角度来看,折扣率计算与利率计算正好相反,所采用的公式是:

$$\text{某数额的现价} = \frac{\text{数额}}{(1+r)^t}$$

在这个公式里,r 代表的是折扣率(例如,.05),t 代表的是年数。一个项目总收益的现价就是将到研究时为止的折扣值的累加。下面是一个例子。

已知某个培训项目能使每个参与者的年收入增加 1 000 美元。将未来 5 年的收入增长按 10% 的折扣率计算成现价。

在 5 年里,总折扣收益等于\$909.09 + \$826.45 + … + \$620.92,或者 3 790.79 美元。这样,在未来 5 年里年增加 1 000 美元的总额的现价并不等于 5 000 美元,而是 3 790.79 美元。如果折扣率为 5%,则现价为 4 329.48 美元。总的来说,在其他条件都一样的情况下,低折扣率计算得到的收益较高。

年度.				
1	2	3	4	5
$\frac{\$1\,000}{(1+.10)^1}$	$\frac{\$1\,000}{(1+.10)^2}$	$\frac{\$1\,000}{(1+.10)^3}$	$\frac{\$1\,000}{(1+.10)^4}$	$\frac{\$1\,000}{(1+.10)^5}$
= \$909.09	= \$826.29	= \$751.32	= \$683.01	= \$620.92

对分析时间的选择,取决于项目自身的性质以及是否适合进行事前或事后

分析。所有其他条件等同的前提下,随时间的延长,项目所获得的收益也越多。

就确定折扣率而言,目前还没有一种权威的方法。其中一种方法就是用资金的机会成本来确定折扣率,就是说,如果把资金投放到另外的项目的话,回报率是多少。不过,必须考虑的是,把资金投放到私营机构(个体机构)和公共机构(准政府机构)的回报率是不一样的,在确定机会成本时,要注意到这一点。同时,还要考虑到投资的时长和风险。

这样,成本—收益分析的结果对于折扣率的选择就变得十分敏感。在实践中,为了解决这个复杂的问题,也为了避免争议,评估者在计算时常常会选择多种折扣率。进而言之,评估者不必使用看上去武断的折扣率,而可以计算项目的**内部回报率**(Internal rate of return),或必须使项目收益与成本对等的折扣率。

与此相关的另一个技术是通胀调整(inflation adjustment),也就是说,在成本—收益分析中,要考虑资产价格随时间的变化。例如,房子和设备的价格会因不同时间段上的货币价值调整而相应地变化。

在前面的论述中,我们谈到的是项目的总收益减去总成本后所得到的**净收益**(Net benefit)。而折扣的必然性意味着净收益应该被更确切地定义为总折扣收益减去总折扣成本。这一总体又通常叫做净回报率(net rate of return)。

将收益转换为货币形式时,许多因素都会导致人们的意见相左。对定价的争论直接牵扯到成本—收益分析对项目效率分析的合法性。

比较成本和收益

成本—收益分析的最后一步就是比较总成本和总收益。如何进行比较,在某种程度上取决于分析目的以及在项目问题上的共识。最直接的比较就是把成本从收益中分离出来。例如,某个项目的成本是185 000美元,收益是300 000美元,那么净收益(或者说,利润,用商业类推法计算出来)就是115 000美元。在通常情况下,使用这种方法的疑问很多,所以,有时人们使用收益与成本的比率,而不是净收益。因为这样的测量通常很难解释,应该尽量避免使用(Mishan, 1988)。

在对成本和收益比较的讨论中,我们已经注意到其与商业决策的相似性,两者是可以类比的。特别是在决定支持哪个项目的时候,一些私营基金采用的是商业投资决策模式。决策者可能会在高风险项目(如,项目的回报率很高,但成功率很低)和低风险项目(例如,项目的回报率很低,但成功率很高)之间进行平衡。这样,基金会、社区、政府部门就希望获得各种可能性和收益估计,来了解"投资风险"。

当然,有时候项目成本会高于收益。专栏11—J列举了一个成本—收益分析的例子,联邦噪音控制项目的负面结果。在这个分析中,控制摩托车、卡车和汽车噪音的成本要大大高于收益。专栏的表格列出了卡车和汽车噪音的控制标准;请注意,从总收益中减去总成本以后,所获得的净收益为负值,收益与成本的比值小于1.0。当然,人们可以对这种收益测量进行质疑,譬如增加每降低1分

贝(dBAs)噪音的价值。然而,根据布罗德(Broder, 1988)的说法,这个分析本身就已经解释了里根政府放弃此项目的原因。

专栏11—1 项目的产生与消亡分析

事情从来如此,政府项目只要启动,就不可能停下来。大多数组织都形成了一种机制,那就是防止组织受到威胁。不过,联邦环境保护委员会(EPA)的噪音控制与治理办公室(ONAC)在里根执政期间却被解散了,终止一项重要的社会项目居然没有引起社会的反对。

尽管终止没有效果的噪音控制项目只是里根政府清理社会项目的一个例子,具有讽刺意义的是,前政府的经济分析却认为这些项目是合理的。更为特殊的是,里根政府还解散了前卡特政府的经济顾问委员会和收入与价格稳定委员会。恰巧是这两个委员会,对几项重要的公共项目进行过经济分析,而里根政府用的也是这些分析。

卡车和汽车噪音控制的成本—收益分析

	控制卡车噪音的规则		控制汽车噪音的规则	
	83 dBAs	80 dBAs	83 dBAs	80 dBAs
收益(a)	1 056	1 571	66.2	188.5
成本(b)	1 241	3 945	358.8	967.3
净收益(a)—(b)	-185	-2 374	-292.6	-778.8
收益—成本比(a)/(b)	.85	.40	.18	.19

注:dBAs = 分贝。除了比例以外,成本和收益的分析或计量单位都是1978年的美元。

资料来源:I. E. Broder, "A Study of the Birth and Death of a Regulatory Agenda: The Case of the EPA Noise Program," *Evaluation Review*, 1988, 12(3): 291-309.

值得注意的是,有时候,产生负值的项目在某种程度上也十分重要,而且应该继续下去。例如,为受重伤人士提供服务是集体的责任,但不是每个完成这项任务的项目都能获得净收益(将用收益减去成本)。不过,在这种情况下,人们也许仍然希望比较各种项目的成本和收益,例如将机构性护理与家庭护理进行比较。

什么时候使用事后成本—收益分析

本章已经讨论了在项目形成中事前分析的重要性。同时还指出,在社会项目领域使用事前分析的概率要远远高于其他。经常出现的情况是,在项目开始实施后,政策决策者和项目主办方才发现,从长远的角度来看,项目成本与收益的比值表明,项目不宜实施。

在事后评估中,是否要进行成本—收益分析,考虑几个方面的问题是相当重要的。在一些评估环境中,进行综合评估在技术上是可行的、有用的,而且具有逻辑性;在其他的评估环境中,使用成本—收益方法也许只能停留在假设推理阶段,而且发挥的作用有限。

对项目进行事后成本—收益分析的最佳前提包括以下几个方面：

- 项目有独立的或单独的经费。这意味着能够从其他活动中分离出项目的成本。
- 项目形成期已过，而且可以确认项目有显著的净效果。
- 项目影响和影响的大小已知，或可以进行有效估计。
- 收益可以转换为货币形式。
- 决策者在考虑备选项目，而不是考虑是否应该继续现有项目。

事后效率估计（成本—收益和成本—绩效分析）应该成为许多影响评估的组成部分。在专栏 11—K 中，讨论了棉纺厂更换机器所产生的大量灰尘。在表格中，维斯库思（Viscusi, 1985）提供了两组数据，结果显示，项目减少了患棉纤维吸入性肺炎的病例数和长期残疾的病例数，等到项目完成，减少的数量正好与估计值吻合。成本数据显示，防止致残的总成本少于 1 500 美元，最为保守的工厂老板知道，与保险公司硬性规定的保险金额相比，项目花费的金额要少。同时，只靠强制执行职业安全和健康管理（OSHA）标准所能减少的棉纤维吸入性肺炎病例数，对工厂老板可能就没有太大影响（如果不进行示范以展示项目的低成本）。

专栏 11—K　棉纺厂粉尘法规：职业安全和健康管理的一个成功案例

20 世纪 70 年代末，职业安全和健康管理委员会（OSHA）在尝试提高纺织行业工人的健康状况上，迈出了坚实的一步——提高了棉纺织工厂的粉尘规范标准。但是，职业安全和健康管理委员会的“棉花纤维粉尘标准”普遍被认为无效，作为美国联邦法院的一个重要决定，备受各方政治力量的争议。然而，有证据显示：这一标准对于工人健康的维护有显著的积极效果，其推行成本也要比最初预期低。例如，关于棉花纤维粉尘环境暴露和疾病发生之间关系的有关数据，以及工伤数据和工人更替方面的证据都显示：患棉纤维吸入性肺炎的风险已经在法规实施之后，有了显著的下降。另一方面，防止“完全致残”的总成本低于 1 500 美元，所以，加强棉花纤维粉尘标准的经济基础很强大。

棉纤维吸入性肺炎患者估计减少量和棉花纤维粉尘标准引进之间的关系如下：

患者类型（患病程度）	每年减少人数（1978—1982）	完全遵守规范下的年均减少量
1/2 级和 1 级	3 517	5 047
1 级以上	1 634	2 349
部分残疾	843	1 210
完全残疾	339	487

资料来源：W. K. Viscusi, “Cotton Dust Regulation: An OSHA Success Story?” *Journal of Policy Analysis and Management*, 1985, (4) 3: 325-343. Copyright© 1985, John Wiley & Sons. Inc.

成本—绩效分析

即使项目目标不是共同的,成本—绩效分析也能让评估者对不同项目的经济效果进行分析。20 世纪 70 年代早期,在社会项目领域,成本—收益分析刚刚起步,然而,在将生育控制与健康、住房或教育项目的成本—收益进行直接比较时,有些评估者仍持有怀疑的态度。正如我们已经注意到的一样,要想对一些重要问题的估计达成一致,根本就不可能(例如,分析生育控制项目的货币价值或健康项目所挽救生命的价值),更不要说结果的比较了。

对多重目标项目而言,可以把成本—绩效分析看作是成本—收益分析的扩展,两者使用的方法和理论是一致的。例如,两种分析方法的假设、测量成本和折扣方法的程序都是一样的。因此,用于成本—收益分析的理论和方法介绍,也可以作为理解成本—绩效分析的基础。

与成本—收益分析相反,成本—绩效分析并不要求把成本和收益简化为一个公分母。相反,需要考虑是,在一定货币成本下,项目获得的绩效。在成本—绩效分析中,可以对具有相似目标的项目进行评估,并比较其成本。这样,人们能够对两个或者是多个旨在降低生育率的项目,或对旨在提高学生学习成绩的不同教育方法,或对各种旨在降低出生婴儿死亡率的项目进行比较。

这样,在考虑达成目标或达成不同层次目标的成本或投入时,成本—绩效分析允许项目之间进行比较,并根据效果分类排队。但是,由于对收益的考量没有一个公分母,所以无法在货币的意义上来断定某个项目的价值。同样,我们也无法确定在不同领域中,在这两个或多个项目之间,哪个效果最好。如果有相同或相似的项目目标以及相同的产出测量的话,所能比较的也只是不同项目的相对效率。在这些分析中,效率指的是单位产出的成本。对于产出相似且无法用货币形式表达的多个项目而言,成本—绩效分析尤其适合。此外,如果已知服务或项目要产生积极结果,成本—绩效就可以用每个项目对象所消耗的成本来表达。识别单位成本,就可以对提供相似服务的不同项目或具有多重服务的不同项目部分或不同次级项目之间的效率进行比较。专栏 11—L 提供了对吸毒者进行美沙酮替代治疗的成本分析。对于评估者而言,他们的特殊兴趣在于,成本—绩效分析可以说明,与标准项目比较,每个培训对象花费的相对成本是多少。当然,就相同项目而言,成本—绩效分析还能揭示每个项目对象花费的成本在 4 个不同的项目点之间的差别。

专栏 11—L 美沙酮替代治疗项目中培训与就业服务的成本分析

先前的评估研究已经表明,给予吸毒者职业和就业咨询,不仅对其就业产生正面的效果,对遏制毒品使用和犯罪也有正面的影响。尽管这些结果令人鼓舞,但由于项目重点的转移或经费压力的增加,许多治疗项目都缩减或取消了职业咨询服务。针对这个问题,"三角研究院(Research

Triangle Institute)"的评估者对四个有就业咨询服务的美沙酮维持项目进行了成本分析，目的是帮助决策者在将来针对吸毒者的项目决策中考虑恢复就业咨询服务的可行性。

在这些项目中，标准治疗是指为吸毒者提供12个月或者更长时间的美沙酮维持、每月进行一次随机尿检、每月都有个人咨询、每个月都有1～4个小组咨询。

这些项目中的培训和就业部分(TEP)包括：职业需求评估，使既有的培训和就业项目适合美沙酮使用者，培训他们并使他们获得工作。每个项目都有一位驻地职业问题专家，主要针对吸毒者和需要了解职业问题的人员提供咨询、提供相关工作信息，并每周与服务对象取得联系。

对标准美沙酮治疗(STD)的随机影响评估结果以及STD与TEP比较的结果表明，美沙酮治疗对象有着很高的失业率并缺乏职业咨询服务，TEP使他们获得了这种服务、得到了培训、减少了他们的短期失业现象。

有了这些积极的产出，那么，重要的实践性问题就是，加入TEP部分会在多大程度上增加项目的成本。为了获得答案，在与标准项目(没有TEP)的成本(四个项目点)进行类比的基础上，评估者考察了TEP的总成本和每个项目对象的成本。主要结果如下表。

分析结果表明，在有TEP的项目中，每个项目对象的干预成本在1 648～2 215美元，分别是STD(没有TEP)的42%～50%。

增加TEP与STD的年总成本和单位(美元)成本比较

	项目A	项目B	项目C	项目D
人员	$38 402	$41 681	$49 762	$50 981
支持职业专家的成本	11 969	14 467	17 053	6 443
差旅	1 211	3 035	2 625	1 870
其他	7 736	14 033	2 619	2 728
总的TEP年均成本	59 318	73 217	72 060	62 022
TEP服务对象	36	38	43	28
TEP每个对象花费的成本	$1 648	$1 927	$1 676	$2 215
STD年总成本	$819 202	$1 552 816	$2 031 698	$1 531 067
STD服务对象	210	400	573	300
STD每个对象花费的成本	$3 901	$3 882	$3 546	$5 104
总的TEP成本/总的STD成本	7.2%	4.7%	3.5%	4.1%
平均TEP成本/平均STD成本	42.2%	49.6%	47.3%	43.4%

资料来源：M. T. French, C. J. Bradley, B. Calingaert, M. L. Dennis, and G. T. Karuntzos, "Cost Analysis of Training and Employment Services in Methadone Treatment," *Evaluation and Program Planning*, 1994, 17(2): 107-120.

由于许多美沙酮维持对象不适合接受培训和就业服务，所以，不可能把TEP

应用于整个 STD 案例。从加入 TEP 以后的项目成本来看,TEP 仅仅使项目总预算提高了 3.5% ~7.2%。同时,分析还显示,正如干预对象单位成本所展示的那样,四个项目给那些同时被提供 TEP 和 STD 服务的对象所带来的效果是不同的。

一些主办方和项目工作人员对效率分析有偏见,因为他们处理的问题主要是"钱",而不是"人"。尽管如此,项目各方用以评定项目实施实整性或维持必要性的方法却并无很大差别。由于地球上的资源是很有限的(尽管这是老生常谈),因此我们要从诸多项目中进行选择,进行资源配置。合适的效率分析能为判断项目的经济潜力或实际成本提供有价值的信息,因此,对于项目的计划、实施和政策决策以及得到项目各方不断的支持而言,效率分析都是十分重要的。

小　结

- 效率分析为考虑项目成本与收益之间的比较提供了分析框架。如果说成本—收益分析是用能相互比较的尺度(货币价值)直接把收益和成本进行比较的话,那么成本—绩效分析就是用单位产出的货币成本来进行表达。
- 效率分析需要使用复杂的技术手段和对相关人员的咨询。作为考核项目产出的一种方式,效率分析直指成本和收益,而且对评估领域也有着巨大的价值。
- 在项目实施的全过程中,从规划、实施到修改,效率分析都十分有用。从目前的情况来看,在社会项目中,使用**事后分析**的概率大于**事前分析**,因为在项目完成之前,对成本和收益貌似合理的估计,常常是缺乏根据的。然而,人们应该更多地使用事前分析,尤其是对那些不仅实施费用极高,而且评估成本也高昂的项目。不同的假设会产生不同的分析,这些分析将要揭示的是在各种假设下,获得预期净收益的不同概率。
- 效率分析使用不同的假设也会相应地产生不同的结果。如果采用不同的会计视角,从单个对象或参与者角度、项目主办方角度或集体(社会)角度来看,都是这样。要选择哪一种视角,则取决于使用分析结果的人和组织,因此涉及政治选择过程。
- 进行成本—收益分析要求知道项目成本和收益的信息,这样的信息必须有质量,并能被转换为一般的测量尺度。将结果或收益进行货币转换的方法有货币测量、市场价值、计量经济估计、参与者估计以及对政策选择的观测。在市场价格不可知或在某些情况下使用市场价格不现实时,就要使用影子或会计价格。
- 在估计成本时,机会成本的概念能帮助进行更为准确的评估,但机会成本的实际应用往往十分复杂,而且充满争议。
- 项目的真实效果包括外部效果和配置效果,在全面的成本—收益分析中,应该将二者都纳入考虑。
- 在成本—收益分析中,成本和收益必须被投射在项目远景中来反映预期变化,以体现项目的长期效果。另外,预期收益和成本必须折算为现价。
- 在项目收益不能转换为货币形式时,可以用成本—绩效分析代替成本—收益分析。成本—绩效分析可以比较有相似目标项目的相对效率,甚至可以用来分析一个项目不同变型之间的相对绩效。

基本概念

会计视角(Accounting perspective):决策中所隐含的、按照商品和服务类别所进行的成本与收益分析。

收益(Benefits):项目积极的产出,往往在成本—收益分析中用货币形式表示,常常与成本—绩效分析中的成本相对应;包括直接与间接收益。

成本(Cost):即投入量,包括直接与间接地用来完成一项任务的投入。

折扣(Discounting):在评价成本与收益时对时间因素的处理方法,也就是说,如果按现价计算成本与收益,就需要选择一定的折扣率和时间框架。

配置性效果(Distributional effects):在普通人群进行资源再配置所产生的项目效率。

事前效率分析(Ex Ante efficiency analysis):项目实施前进行的效率分析,通常是项目计划的一部分,用来估计相对于成本而言的净效果。

事后效率分析(Ex Post efficiency analysis):项目实施后的效率分析,为的是了解项目净效果。

内部回报率(Internal rate of return):使项目总收益与项目总成本相等的折扣率计算值。

净收益(Net benefits):总收益减去总成本的剩余额,也被称为净回报率。

机会成本(Opportunity costs):由于项目干预的影响而造成机会流失所损失的价值。

次生效果/外部性(Secondary effects/Externalities):将项目成本花费在非目标人群身上所造成的效果。

影子价格(Shadow price):并非按照精确市场价格所估计的产品与服务价值,也指由规则约束或外部性所导致的不恰当市场价格,也叫做会计价格。

12 评估的社会背景

在本章,我们主要关注的是实施评估活动的社会和政治背景。从一开始我们就已经表明,评估并不只是调查程序的简单应用。评估研究是一项有目的的活动,其效用在于影响政策发展、设计和完成社会干预、改进社会项目的管理。从广义上讲,评估是一项政治性活动。

当然,评估研究者还可以从评估活动中获得某种内在的情感回报。他们可能会因为尽善尽美地完成了一项技术性工作而自鸣得意,就像那些将画作束之高阁、从不示人的画家,像那些对手稿秘而不宣的诗人。但事实上,评估并不完全如此。评估是一种现实生活中的行为,至关重要的是该领域内同行的褒贬和对政策、项目及实践(在短期或长期,对人们生活环境的改善)的修正作用。

评估实践者形形色色，他们在学术观点、意识形态和政治方向以及经济、职业取向等方面各不相同。尽管具有这些差异，但几乎所有评估者在评估工作目的性这一点上，都有一致的观点。大多数人认为，应用性工作的基本原则是影响社会某阶层主要人群的行为和思想（这部分人群可以直接影响社会变化），或在政策和行为中使用评估者提供的调查结果和结论。

当然，我们已经跨入21世纪，与20世纪70年代末《评估：方法与技术》第1版出版时相比，不难想象，今天，在评估界中出现了更为复杂的情况，不仅在技术方面，而且评估研究在政策和社会项目领域内的地位也发生了变化（如果要纵观该领域的发展与变化，可参看 Chelimsky and Shadish，1997；Haveman，1987；Shadish，Cook，and Leviton，1991。对于评估演变的不同观点参看 Guba and Lincoln，1989）。与此同时，在方法问题、评估者的培训和评估行为的组织化管理等方面都存在着争论。而且，该领域仍面临对于评估者社会责任感的政治和意识形态问题的争论，在对研究发现选择最有效的传播途径上，尚存在分歧，在评估效用最大化的最佳策略方面亦有观点差异。由于过去数十年的经验，评估者对他们努力的作用评价变得更加保守，他们开始认识到社会政策不能仅仅依靠评估。甚至评估事业最有力的拥趸者也承认，评估对于社会政策可能产生的作用受制于承担评估研究者和他们客户的能力高低以及个体兴趣范围、受制于工作风格和组织安排的多样性、受制于在一切计划好的社会变化实施过程中不可避免的政治和经济方面的约束。至关重要的是，在一个民主社会中，社会变化不能仅靠专家意志决定，更应该是一个在权衡多方利益之后得到的结果。

另外，大多数评估者相信社会项目可以改善人类生活条件，然而他们发现很多社会项目并没有带来显著的社会改良，有的甚至根本无效，评估者因此感到失望。我们已经知道，项目设计有效并得到完全实施是非常困难的。对于很多人而言，这给他们带来的不是令人振奋的经验而是坏消息。

因此，评估者体验了挫败、无力感和群体自尊心的缺失，他们做出的努力往往缺乏信心和方向。他们的反应也是相同的：过度内省、齐心协力把过失推给别人、对社会和人类事态阴暗面夸大其词地评述，特别是强调发展和完成有效社会干预的无用性。一些社会评论者甚至根据盛传的评估失败案例，断言当前评估实践经常不能有效地识别成功的社会项目（Schorr，1997）。

显然，无论评估规划如何得当、操作如何精密，仅仅依靠评估本身，并不能根除我们人类和社会的问题。但是，评估工作可以把我们导向预期的方向，这一作用是应该承认的。有相当多的证据表明，评估结果的确经常影响到政策制定、项目计划和执行以及社会项目的管理方式，只是这种影响可能在短期内发挥作用，也可能在一个长期的过程中体现出来。

评估活动在逻辑上被归入"应用"社会研究这个大范畴之下。虽然区分"基础"或"学术"研究与应用研究的界线通常不是十分清晰，然而二者仍然存在着本质上的差别（Freeman and Rossi，1984）。我们在前面已经论及了二者的一些差别，譬如，我们指出评估应该得到必要的引导，使其完善到可以解答学术研究提

出的问题。这种实用主义本位与进行基础研究的学者形成对比,基础研究的特点是努力找到完成研究的“最佳”方法。当然,基础研究也要受制于现有研究手段,因此折衷方案经常是必要的。

理解应用社会研究与基础社会研究的另外三大区别十分重要。首先,基础研究的特点是为了满足研究者智力上的求知欲,有助于研究者及其同行在感兴趣领域的知识发展。基础研究通常直接导向学科主要关心的问题。与之相对,应用研究的开展是为了解决实际问题。在评估领域中,推动工作进行的动力经常不是来自于评估者,而是关心某一特定社会问题的个体或集团。因此,评估者理解评估领域中的“社会生态学”尤为重要。这是本章开始讨论的第一个重要话题。

其次,基础研究者一般受单一的学科思想训练,并在职业生涯中遵循这一种思想。他们的特点是使用有限的方法和程序,在一个有限的独立知识领域中进行一项又一项的研究。举例而言,一个经济学家会将医疗费用的研究作为他的专门技术领域,并且一直采用计量经济学的模型进行研究。同样,一名社会学家可能主要采用参与观察方法,大半生从事教育职业生涯研究。与此相反,评估研究者有时候会从一个项目领域转向另外一个领域,去面对不同的问题,他们需要通晓一套研究方法和多个独立研究领域。例如,一位研究者就认为,社会项目评估的内容可能涉及营养学、预防犯罪、自然灾害影响、对儿童的虐待和忽视、标准化舆论以及教育的多层次性等各个方面。评估利用的方法也是多样的,它可以是随机实验,也可以是大规模的截面研究,或者是对文献的数据分析。一些评估者精通一个或几个项目领域,他们多方面的详尽知识与他们的评估专业知识是紧密结合的。评估者经常要面对变动的课题领域,这带来了一系列问题,如与基础研究者不同的评估研究者的训练、视野和理论观点等,概括说来,就是评估的职业化(Shadish and Reichardt, 1987)。评估职业化是本章的第二个主要关注点。

再次,虽然伦理关系对于基础研究和应用研究都很重要,然而在实际工作中显得更为迫切,具有更大的社会意义。如果一个基础研究者违反了他的职业准则,会给所从事的学科带来损害。但是如果一个实地调查员逾规,则可能影响到评估的项目、目标人群,甚至整个社会。因此,本章接下来将主要关注的是在应用研究中的伦理问题。

第四点,基础研究和应用研究在受众和判定其有效性的标准方面也有较大差异。基础研究者最关心的是同行们对其研究的意见;研究是否有用要看论文能否在权威期刊发表,能在多大程度上影响到其他人的研究。应用研究依靠的则是自我判定和研究主办方的判定,主要看它们能在多大程度上促进政策和项目的发展和完成,其最终目的是解决社会问题。评估结果的利用以及使其效用最大化的途径,将是我们最后进行讨论的主题。

此外,我们承认,每项评估都有各自的特点,需要针对所涉及的问题,采用特殊的解决方法。正因为每项评估均具有独特性,才造成了为评估行为提供细致“准则”与具体引导的困难。不过如今,这个领域已经发展得比较成熟,可以为我

们提供比较合理的评估技术现状评论、评估实施的一般性指导和建议。本章就是结合了我们的经验以及与评估相关的人际关系、政治和结构等方面研究者的成果。

评估的社会生态学

评估能否被人们利用，取决于评估者是否认识到决定评估效用的关键因素是评估进行的社会和政治环境。因此，为了成功地完成一项评估活动，评估者需要不断地对他们的工作环境进行社会生态学的评估。

有时候，一项评估的推动和支持力量来自于最高决策层：由国会或者联邦机构授权对社会革新项目进行评估，如健康与人类服务部对收入维持项目中属于政府的项目进行评估（Gueron and Pauly, 1991）；或大型基金会的会长主张对该基金会的主要社会项目进行评估，如约翰逊基金会的住房供给资助项目（Rog et al., 1995）。也有评估活动是应各种生产机构的经理和管理者的要求进行的，其重点是针对这些机构和项目方的管理事务（Oman and Chitwood, 1984）。还有一些评估是由社区中的个体或群体推动的，他们与某一社会问题有着直接的利害关系，并准备或者正在处理问题。

无论动力是什么，评估者的工作总是在一个真实世界的背景下展开，这一世界充满各不相同的甚至经常是相互矛盾的利益群体。在这种关系中，必须注意两个作为评估背景的重要因素：一个是多项目方的存在，另一个则是评估通常是政治程序的一部分。

多项目方

在研究中，评估研究者通常会发现，在他们的工作和研究结果中，存在着代表不同利益的群体和个体。这些项目方在评估工作的适用性及评估结果将影响谁的利益等方面观点相左，有时甚至完全对立。为了有效地开展工作并解决问题，评估者必须对自己和项目各方的关系、不同项目方之间的关系有透彻理解。

项目方的范围

不管是在评估者开展工作伊始，还是在后来他们对结果的反馈中，要理解这些关系，必须先搞清楚哪些项目方可以直接或间接影响评估工作的有效性。不管是在一个学校、一家医院或一个社会机构中独立工作，还是那些在大型的组织化调查中心、联邦和政府部门或者基金会中联合工作，评估者们都要面对这个问题。从理论上说，任何一个关心社会生活环境改善成效的公民，都会关注社会项目评估及其结果。当然，在实践中，任何评估活动的范围要窄得多，只包含在项目中有直接可见利益的人群。在项目方群体的内部，不同的项目方对评估结果的意义和重要性会有不同的观点。这种观点差异为项目各方之间以及项目方与

评估者之间的可能冲突埋下了隐患。不管一项评估结果如何,对一部分人来说是好结果,对另外一部分人来说则是坏结果。

评估就是做出判断,而评估的过程就是为判断提供经验证据。区分进行判断与提供判断可依据的证据二者之间的差别,在理论上是有用的、能辨明的。但是在实践中却通常很难做到。无论评估者对社会项目有效性的判断是通过怎样严密的研究设计和精确的数据分析产生的,一些项目方仍有可能认为评估结果是武断的,并做出相应的反应。

我们对如何组织与激活评估各方所知甚少,不太清楚项目各方的利益所在,也不清楚各方利益是如何确立的;同样,不了解他们如何就某一评估结果而发生利益表达。也许,唯一可以断定的是——最有可能自始至终关心评估活动的参与者就是评估活动的主办方和项目经理及职员。当然,这类人群通常与项目关系最为密切,因为他们的活动将在评估报告中得到最直接而清楚的评判。

项目的受益者或称为项目目标人群的反应尤其关键。在许多案例中,受益者与评估结果的关系最为密切,但他们经常也是最无意发表意见的人群。目标群体总是无组织的、分散的;他们往往没有受过什么教育,在政治联合方面也没有经验。他们有时候甚至不愿意表明自己的身份。目标群体即便在评估过程中表达自己的意见,也往往是通过那些希望成为其代表的组织机构。举例而言,无家可归者很少能够在减轻他们生活痛苦的项目讨论中发表意见,所以,全国无家可归者联合会,一个主要成员并非无家可归者的组织,经常在处理无家可归者的政策讨论中充当他们的代言人。

多项目方的后果

多项目方的存在可以导致两个重要的后果。其一,评估者必须承认,即便他们的努力可以影响到决策和行为结果,但也仅是影响这个复杂过程的因素之一而已。其二,项目各方之间的利益冲突常常会给评估者带来紧张。这种紧张一部分可以通过预期和计划来削减或使其最小化。另一部分则由某些特权造成,必须在特殊的基础上加以解决或者只能视其存在为理所应当。

评估中的多项目方主要从三个方面给评估者带来紧张。首先,评估者经常无法确定在评估设计中应该采纳哪一方的观点。整个社会、政府机构、项目职员、委托方、一个或多个项目方,究竟哪一方的观点是合适的? 对于有些评估者,尤其是想为运行良好的社会项目提供帮助和建议的评估者,其首要的评判者(读者群)是项目职员。对于承担立法机构委托任务的评估者,首要的评判者则是社会、政府或国家。

在一项评估中,采用哪一方的观点或立场不等同于接受哪一方的偏见,这一点很重要。所采用的观点关系到项目的目标界定和决定应该考虑项目方的哪些(关注)利益。与此相反,评估中的偏见通常意味着歪曲评估设计,以获得符合某些项目方要求的不公正结果。每一项评估都必须在一套观点立场的指导下进行,但一项符合职业伦理的评估应该努力避免在设计或分析中存在有偏见的评

估结果。

一些评估学者尤为强调评估研究行为应该受一些特殊观点、立场的支配。"关注利用的评估(Utilization-focused evaluation)"方法(Patton, 1997)主张,评估设计应当反映"主要使用者"的利益,为可能在特定案例中起决定作用的人制订特殊的评估方法。"授权评估(Empowerment evaluation)"的倡导者(Fetterman, Kaftarian and Wandersman, 1996)认为,评估的目的在于授权给被忽略的群体,如穷人、少数派,应当采纳他们的意见,吸收他们参与评估的设计和分析。必须指出,以上两种方法在观念上都不算是偏见。评估者到底要持什么样的观点,我们的立场并不明朗。在第 11 章,我们讨论了进行有效的分析不同观点。我们认为,没有哪一种观点是一定恰当的,不同的观点都可能是合理的。从客户或者目标群体出发的观点并不比从项目本身或者发起项目的政府机构出发的观点更具合法性。评估者的责任并不是从许多观点、立场中找出一个合法的,而更应当是在清楚认识到其他观点、立场存在的同时,明确一项特殊的评估活动应当遵循的那一种观点立场。举例而言,在对评估结果的报告中,评估者应当说明该项评估是以项目管理者的意见为主导进行的,同时也应承认存在着社会整体和目标群体在这许多问题上的观点分歧。

在一些评估中,对项目可能有不同的观点。譬如说,对于项目对象而言,收入维持项目意味着政府要减少满足基本需求的资金投入;从联邦议会的立场来看,项目的主要目的是促进受资助者早日摆脱资助而获得自立,低水平的扶助不失为一个可取的刺激因素。收入维持项目的受助者认为,成功的项目应该是能够提供给受益人足以应付基本消费的资金;而议员们则认为,庞大的资金扶持会助长福利依赖。

第二,评估者必须认识到,当评估结果显示一些社会政策或项目的价值并不像评估主办方原先鼓吹的那样时,评估主办方可以向评估者提出质疑。虽然评估者一般可以预计到其他项目方的消极反应,但是对评估主办方发现结果与预期相反时候的反应,他们往往措手不及。这种时刻,评估者就处于一个十分困难的境地。失去了主办方的支持,评估者就将直接面对其他项目方的攻击,而他们原本期望主办方可以为他们抵挡这些攻击。必须考虑到的一点是:主办方是应付评估研究之外附加事务的最佳人选,也是评估项目所需资金的主要提供者。关于这个问题的例证可以阅读专栏 12—A,其中呈现的就是芝加哥无家可归者调查结果受到主办方尖锐质疑的案例。(关于同样事件的截然不同观点,请看 Hoch, 1990)芝加哥项目方的反应并不具有普遍意义,因为我们仍可以举出很多不受欢迎的评估结果最后却被接受、甚至被付诸实践的例子。显然,对于同一项评估结果的社会反应还会随着时间演进而发生改变。譬如,几年之后,罗斯(Rossi, 1989)对于芝加哥无家可归者的评估就受到人们的普遍称赞,认为这是一项整个 20 世纪 80 年代中最成功的、关于无家可归者的描述性研究。

第三,因为存在与各项目方的沟通困难,可能会由此形成误解。与基础性的社会科学学科相比,评估的词汇表并不太复杂深奥。但是,这仍使外行受众比较

难懂或难以接受。举个具体的例子,"随机"概念对于影响评估非常重要。从技术上说,这个概念是精确的、不含贬义的,正如第7章已证明的那样。然而在外行人的词汇中,随机却含有偶然的、疏忽的、无目的的、巧遇的等含义,是贬义的。对于评估研究者而言,主张将调查目标随机分配为实验组和控制组是严密的、区分严格的方法,然而对于外行们的意义却是截然相反的。因此,如果评估者不严格界定"随机"的含义就随便使用的话,会具有一定的危险性。

如果要求评估者收集到与评估有关的所有详尽的、各类人群的信息,实在是苛求。信息沟通对于评估程序的理解和评估结果的利用仍然是一个障碍。因此,评估者应当预料到在与不同项目方交流中可能出现障碍。我们将在本章后面的篇幅中对此做详细讨论。

专栏12—1 相反结论的后果

在20世纪80年代中期,约翰逊基金会和教会基金组织授权马萨诸塞大学社会与人口统计学院提供统计无家可归者详细情况的可行方案。这两个基金组织刚刚启动了一个为无家可归者建造诊疗中心的项目,需要无家可归者的准确数据来评估诊疗中心如何更好地提供服务。

我们关于芝加哥无家可归者数量的调查结果很快成为争议的焦点。对芝加哥无家可归者的统计工作一向是由专业人士组成的芝加哥无家可归者联盟和市长委员会来完成的,且一直被媒体和公众认为是对该市无家可归者评估的权威机构。他们的观点基本上影响了大众的普遍看法。在他们的调查报告中显示,芝加哥有20 000到25 000名无家可归者,这个数据被广泛引用。

调查初期,芝加哥无家可归者联盟对我们的研究持中立态度。我们对联盟详细说明了这项研究的目的和资金来源。可当我们寻求他们的帮助,特别是向庇护所管理员要求采访无家可归者住户时,他们不置可否,只是对我们的研究表示怀疑,特别是在标准上界定无家可归者的,他们认为标准应该更为宽泛,以便把居无定所的人、全家挤住在一起的人以及租用棚户的人都包括进来。

当第一阶段调查数据出来后,我们都被震惊了,我们所得出的无家可归者数字要比联盟所使用的数字小很多,只有2 344名,而不是20 000到25 000名。由于我们预计了一个比这个数字大得多的结果,也许我们的街道样本太小,不能够获得足够精确的结果。我们开始怀疑是否在抽样设计或实施过程中犯了严重错误。让我们更加困惑的是,表示过支持我们大部分计划的这两个机构也开始怀疑我们的研究,同时,那些站在无家可归者一边的人也趁机直接提出抱怨。更麻烦的是,第一阶段研究耗尽了主办方提供的全部资金,而这原本是打算用来维持一年中三个阶段研究的。在仔细检验了第一阶段研究成果后,我们确信数字的大幅度减小是正确的,但如果能够重复该研究,就能使局外人也相信这一结果。我们成功地使主办方相信并提供更多资金进行第二次调查。这次调查比第一次有了更大的抽样框,街道抽样补充了一些可能含有大量无家可归者的典型区域(公共汽车、电梯、地铁、医院候诊室等),来检验上次调查采用的午夜时点是否错过了大量只是在夜幕刚刚降临时在街上游荡而在我们调查开始前找到住所的无家可归者。

第二阶段的统计数据显示:芝加哥午夜无家可归者数量的估测值为2 020,标准误为275。第二阶段的工作大大提高了评估的精确程度,但却仍旧与联盟提供的那个数据相差甚远。用我们的数据同样可以估测那些因为找到暂时住所或在监狱、医院而被我们的调查遗漏的无家可归者。另外,我们还估测了和父母一起的无家可归儿童数量(在我们的调查中没有发现无家可归的儿童)。把这些数字都加上后,我们得到了一个新的总数:2 722名。这个最终结果仍然和联盟的20 000到25 000

名的数据相差很远。

这份最终报告在同一时间被送到芝加哥的报纸、电视台和相关政党手中,也还是有几份落到联盟的手中。芝加哥的所有报纸都对这个报告大做文章,紧接着的第二天,就有联盟成员对报告公开指责。不管我们怎么努力想把注意力引向我们的研究发现,报纸还是用大字标题强调我们的统计数字。从联盟传来的批评很刺耳,他们指责我们的研究是对无家可归者的极大损害,认为我们试图通过简单的(甚至说是蓄意的)数字估测来转移公众对无家可归者现状的注意力。联盟还暗示我们的报告是在伊利诺斯州公共援助部的授意下完成的,存在技术性的缺陷,认为我们界定无家可归者的标准忽略了数以千计、与朋友和亲戚同住或住在不合标准房子里或每晚都要寻找新住处的人。

在应邀到市长委员会进行研究陈述的时候,我们陷入一大堆指责和批评之中。这其中既有纯技术性的,也有来自已经向政府保障无家可归者工程预售大量产品的企业。但争议的焦点在于:我们的报告给了州和地方官员一个忽视和打发这个问题的借口,从而严重损害了芝加哥无家可归者的利益(事实上,政府伊利诺斯州公共援助部门已经保证将加强在这方面的投入)。那两个小时是我自二战在军队接受训练以来所受到的最长时间的个体攻击。尽管看起来使人厌烦,但我必须为了保护我们经过仔细负责调查所得出的结论而去同凭经验主义的、想象的观点作斗争。

几乎是一夜之间,我就成了无家可归项目领域最不受欢迎的角色。当我被约翰逊基金会邀请,去参加一个在洛杉矶举行的会议上作演讲时(这个会议的主要参加者来自这个基金会所赞助的各个无家可归者诊疗中心),除了一些业外人士,在场的人没有一个愿意和我交谈。我变成了会场上的一个透明人,每个人都刻意地躲开我。

资料来源:Peter H. Rossi, “No Good Applied Research Goes Unpunished!” *Social Science and Modern Society*, 1987, 25(1): 74-79.

评估结果的传播

评估结果若要被使用,必须向主要项目方和大众传播,并且得到他们的理解。对于我们的目的而言,传播是指通过一系列活动使一定范围内的相关听众群知晓评估结果。

传播是评估研究者的一个明确责任。一个评估如果不能被相关的听众群了解,就很容易被人忽略。因此,评估者必须认真撰写报告,为确保评估结果能够传送到主要资助人那里做好准备。

很显然,调查结果必须通过易于理解的途径传播到项目各方。值得一提的是,评估技术组一般要给评估主办者提供一份技术报告,对评估设计、数据搜集方法、分析过程、结论、深入调查的意见以及对项目的建议(就测量或影响评估而言)等方面做出详细而完整的(未必如实的)描述,同时对这些数据和分析的局限性进行说明。技术报告通常只供同行阅读。那些项目方是很少认真阅读的,他们中的许多人不习惯阅读长篇报告,也没有时间去读,或者无法理解这些东西。

正因为如此,每位评估者都必须学会成为“二次传播者”。所谓**二次传播**(Secondary dissemination)是指为迎合项目方的需要而进行的对评估结果和有关建议的传播;与此相对,**初步传播**(Primary dissemination)在大多数时候是指技术报告。二次传播有很多种不同形式,可以是技术报告的缩略本(经常被称为摘

要),也可以是由评估研究组或评估主办者定期发布的更复杂版本的专门报告,可以是备忘录,或者是带幻灯片的口头汇报,有时候甚至是电影和录像。

二次传播的目标很简单:以那些著名的"有智慧的外行人"能够理解的方式传播评估结果,"万事通"这种人有时候和北美野人一样,叫人捉摸不透。评估业内的大多数人并不了解,对二次传播文件进行必要的修饰是一门艺术,因为在非学术训练中很难接触到这方面的知识。二次传播的重要策略是:找到公布研究结果的恰当方式,使用的语言和表达形式可以被那些有智慧的、但没有学过专门术语和规则的听众所理解。对语言的要求是尽可能避免使用深奥的专业术语;对表达形式的要求是二次传播的文字资料应该简洁短小,不要令人望而生畏(对这个过程的有用建议可以参考 Torres, Preskill and Piontek, 1996)。如果评估者没有能力将评估结果进行传播、扩大评估效用的话(我们几乎没有人可以做到),那么授权给专家来进行这项工作是比较合理的。归根结底,正如我们所强调的,评估是有目的的活动,评估结果如果不被使用,就没有任何意义。

作为政治过程的评估

全书中,我们强调在项目发展和运作的每一步骤中,评估结果对于决策过程都有作用。在项目设计的最初阶段,评估可以提供社会问题的基本信息和描述,据此可以设计对应的、合适的服务。如果早期项目已经得到检验,对试点项目的评估可以为项目完全实施后可能出现的最终效果提供估测。在项目实施过程中,评估可以提供有关责任问题的重要知识。但这不等于说原则上有用的东西可以自动被理解、接受和使用。在任何时期,评估都只是政治过程的一个要素。这就是评估应该充当的角色:在民主社会中,有重要社会意义的决议应当通过政治过程来确定。

在一些案例中,主办方对评估寄予厚望,认为评估结果将直接影响项目是否被继续、修改或者终止。在这些案例中,评估者承担着迅速提供相关信息以便进行决策的压力。简而言之,评估者面对的是善于接纳意见的听众群。在另外一些情况下,评估者却在完成评估后发现,主办方对于评估结果反应迟钝。更令人不安的是,有时候项目在没有参考评估者得来不易的评估信息的情况下,就被下令继续、修改或者终止。

虽然在这种环境中评估者会感到自己的努力不过是徒劳,但是必须记住:决策过程实际上是一个很复杂的过程,评估结果仅仅是影响决策的一个因素。早在 1915 年发生的纽约"加里计划(the Gary plan)"评估争议就证明了这一观点(参见专栏 12—B)。在人类服务项目中,可能涉及很多团体,包括项目主办方、管理者和操作者,有时候还有项目的参与者。他们常常与项目的继续进行有重大关系,而且那些往往令人无法忍受但热心的主张,甚至比冷静客观的评估结果更为重要。而且,尽管评估的结果是一个立场鲜明的结论,而按照美国传统的政治过程得到的结果,却是各方利益的平衡。

在必须权衡、评估、平衡互相对立的观点和不同支持者利益的任何政治体系

中，评估者充当的是专业证人角色，他们证明项目发挥效力的程度，支持以事实为基础的信息。决策者和其他利益群体的评判比愚昧的观点或貌似正确的猜想的证明力要强得多，但他们（不是指评估者）必须要形成一个决策。还有一些必须考虑到的另外问题。

想象一下，如果评估者被赋予了在政治决策过程中的否决权，这等于是剥夺了决策者的特权，将会出现怎样的情形。在这样的形势下，评估者将成为"无冕之王"，他们对某个项目的决断将不再顾及其他任何团体。

一句话，评估的合适角色应该是为政治过程提供尽可能最佳的、针对评估问题的知识，而不是试图取代这一过程。在专栏 12—C 中，笔者摘引了作为现代评估理论奠基人之一的坎贝尔（Cambell）的一篇文章，解释了将评估者作为"实验性社会"服务者的观点。

专栏 12—B　政治与评估

这是关于新的学校组织形式在第一次世界大战期间如何引入纽约的一个例子。这个被称为"加里"的计划把学校按照新型工厂的模式来建设。就像把孩子们放在移动架上，在平台上从一个学科向另一个学科移动。接下来要介绍的是，评估结果的影响如何渗透到新学校董事会与既有学校管理者的政治斗争中。

加里计划是被一个有改革精神的、市长任命的新学校董事会引入校园的。起初只是处于实验阶段。一个叫马克斯韦尔（Maxwell）的校长非常不满这种对他职业的干涉，对市长的意图也表示怀疑。他在加里计划试点时就表达了对此的看法："当我某天参观学校时，唯一所能见到的就是许多孩子在埋头读书。"尽管这位校长表达了他的不满，但加里体系还是扩展到了布隆克斯（Bronx）的 12 所学校，而且还有进一步扩展的计划。一名学校董事会成员呼吁在扩展这个计划之前多做一些调查研究。

1915 年夏天，加里计划开始在纽约州各学校实施的同时，马克斯韦尔校长要求做一次关于该计划实施情况的评估研究。这项工作被交给了白金汉（Buckingham），一位在纽约市学校研究部门工作的教育心理学家，也是一位在学术成就检验方面的开拓先锋。他使用了自己最新的学术成就检验方法去比较 2 所加里式学校、6 所其他比较组学校，以及 8 所传统式学校。结果是：传统式学校平均成绩最好，而加里式学校成绩最差。

白金汉的报告是对加里体系拥护者的极大批评，批评了他们的不成熟言论。报告一出现就在报纸上和学术杂志上引起了一场反驳的风暴。纽德（Nudd），作为公众教育协会的负责人就白金汉的报告写了一个详细的批评，发表在《纽约环球报》、《纽约时报》、《学校与社会》以及《教育期刊》上。纽德认为白金汉进行测试的时候，加里计划刚刚在一所学校里实施了 4 个月，在其他学校还不到三星期。他强调许多必要的设备还没能提供，而且加里学校的日常工作被大量想要看看这项计划的参观者所干扰。在一个详细的学校与学校的比较中，纽德表示，在其中的一个加里学校，90% 的学生来自于移民家庭，意大利语是他们的母语，而其他一些用来比较的学校的学生则大多来自土生土长的中产阶级家庭。而且，其中一所加里学校的学生成绩与其他学校相比非常出色，但当分数被另一所学校平均后，整体成绩就落后了。

白金汉没有在这场有些失控的争论中做出任何回应，但他指出：他并非仅仅对 2 所、6 所或是 8 所学校进行检测，而是在 11 000 名儿童中比较衡量，因此他的研究是对加里计划的一个真实检验。

他证实了在他刚刚开始调查时,加里计划就突然推向了纽约市的所有学校。正如上面所提到的,在向全纽约市的学校推广这个计划并增加在这个方面的预算时,受到了来自市长办公室的压力。教育委员会的主席发现,在会议的争论中引入纽德对白金汉报告的批评会非常有利。而马克斯韦尔校长直到一年半以后还在继续引用白金汉的研究成果来作为对抗加里计划的证据。

资料来源:A. Levine and M. Levine, "The Social Context of Evaluation Research: A Case Study," *Evaluation Quarterly*, 1977, 1(4): 515-542.

专栏 12—(作为现实社会服务者的社会科学家

我们的社会仍将使用占据优势的、非科学的政治过程来决定创新改良项目的实施。提高社会科学在项目决策中的地位是否更好,在这里并不作讨论。问题的关键在于:帮助社会判断创新能否在没有负面影响的同时,帮助改变达到了预定目标的社会科学家的被动角色。实验社会方法论学者的任务不是告诉我们将要做什么,而是已经做了什么。根据我们对社会科学的了解,得到应用的主要是调查方法论,远甚于描述性理论,应用的目的在于可以从中获得比由政治过程决定的创新更多的知识。看起来这一目的与目前作为政府顾问的经济学家、国际关系教授、外事专家、政治学家、研究权利和民族关系的社会学家、研究儿童发展和学习问题的心理学家等角色差异很大。政府提出一个问题,学者用本学科的科学数据加以解决。在这样的过程中,学者在意见得到遵从的时候,就陷入了被盲目拥护的泥潭,失去了对问题刨根究底的兴趣。一个人如果已经得到了肯定,就不可能再发现他的理论到底有多大作用。社会科学家应该比自然科学家更谦虚,更应该实事求是。也许我想要说的是,社会科学家应该表明他们的意见是由精确的测量获得的,并且需要在实践中得到检验,而不是来自于那些夸夸其谈的伪科学。

资料来源:Donald T. Campbell, "Methods for the Experimenting Society," *Evaluation Practice*, 1991, 12(3):228-229.

政治时机和评估时机

与学术性社会调查不同,在评估中还存在着其他两种压力(由评估者从事的、包含不同利益群体的政治过程所决定的):一个是评估与政策的密切关系和重要性(这个问题我们即将开始讨论);另一个是政治时机与评估时机的区分。

评估需要在一定时间内进行,尤其是在评估项目效果的时候。通常是研究设计越严密、质量越高,操作评估所花费的时间就越长。对重要创新性项目影响的大规模测量、社会调查和报告,需要 4 ~ 8 年的时间。政治环境和项目环境常常发生剧变。决策者和项目主办者通常没有耐心知道一个项目是否可以达到预定目标,他们的时间维度常常是月而不是年。

因此,评估者常常迫于压力,缩短最佳方案所要求的时间来完成评估工作,发布一个初步的结果。有时要求评估者提供项目有效性的估计,而评估者强调如果没有更成熟的产出,做出这样的估计很可能无效。举例说,1997 年和 1998 年,大众传媒和立法委员询问评估者,1996 年的个体职责与工作机会法案中提出的福利改革效用如何。其实那时,在几乎各个州,这项改革还没有详细落实,只

是刚刚开始。而真正的效果要在法案实施 3 ~4 年以后才能显现。对评估所能提供的证据的需要迫切,以至于媒体相信小道消息、轶事,甚至是无根据的猜想。

另外,与启动评估相关的、在主办者工作的组织机构内进行的计划和程序工作,使得及时完成研究十分困难。在大多数案例中,程序必须通过很多个主要项目方的层层审批。结果,光是得到授权进行评估,就要花去大量时间,这还不包括实施和完成评估所需的时间。虽然政府和私营机构的主办者都试图改进机制以加速项目和实施评估的进程,但是这些努力受到了官僚作风、契约性法律程序、对评估的问题和设计达成一致等因素的阻碍。

我们仍不知道如何才能减少评估者与决策者在进度表之间的差异。有时,还需要我们判断的是:到底是了解一些信息好,还是根本不了解一点信息更好。至少,评估者应该按项目方(尤其是评估的主办者)的要求来预定时间,避免做出不合理的时间安排,这很重要。显然,如果在评估结束之前,需要的信息都已经搜集完整,就没有必要进行长期研究。

一个策略性的方法就是限制那些针对将来不可能大规模实施的社会项目所开展的、技术上复杂的评估,在项目被纳入决策团体议事日程之前,代之以小规模随机实验方法开展针对新项目的评估。

评估者的另一个策略是事先预测项目和政策活动发展的方向,而不是被动地接下在规定时间内根本不可能完成的工作。一个重要建议是,建立独立评估机构,致力于检验项目的探索性工作,这些项目有可能在某天被列为评估对象。可以建立评估中心来不间断地评估为解决社会问题而提出的可供选择的社会项目的价值,这些社会问题有的一直受到关注的,有的则是有极大可能在若干年后出现。虽然这个建议听起来前景美好(特别是对专业评估者来说),但是不是能够精确预测影响下个 10 年的社会问题实在是个问题。在对新项目进行事前评估方面,最成功的例子也许是美国审计总署(GAO)的项目评估部和方法部联合进行的一个事前评估。正如专栏 12—D 中切里姆斯基(Chelimsky, 1987)所描述的,部门的事前活动可以为社会规则的制定做出重要贡献(Chelimsky, 1991,对于实用社会调查如何与政策制定相结合的观点)。然而,从目前形势看来,我们相信:政治时机和评估时机差异带来的冲突仍将是一个大问题,这一冲突将伴随评估作为决策者和项目经理的有用工具而存在。

专栏 12—D 在对新项目分析中实施评估活动

我们中的许多人在溯本追源的研究中花费了大量时间,这仍然是评估研究中不可缺少的部分。国会要求我们和行政部门做这些研究,这的确是必要的。但在政治性调查中插入这种研究就不是一件容易的事情了,有时候甚至很不顺利。比较而言,在一个项目开始前,评估者往往在改进评估理论、鉴别调查对象和寻找最佳时机与方法等方面产生了很大的影响。但是,如果新项目进行的节奏过快,就往往会带来一些困难……如果评估结果需要很快投入应用的话,那么评估的步调常常会变得不可控制,以至于无法获得评估工作所需要的足够时间。

在 GAO,我们实行了一套被我称之为“评估计划评论”的方法,这套方法在新项目的程式化方面

有很大的作用。我们在"青春期怀孕"项目上做了初步的尝试。事实上,这套方法就是搜集以往的信息以及相关的研究成果,并给这个新项目结构提供经验性的帮助。参议员查非(Chaffee)让我们审核了他提出的法案;我们花费了四个月的时间去做这项工作,这不仅仅是我们,也是公众关于立法方面的一个大的成功。从更高的政治性角度来看,预先了解项目可能的工作情况将可以使其变成有价值的公众服务,或可以对考虑不周的项目计划进行弥补。事实上,有些问题是决策者不愿意让评估者了解的,因为这可能会使问题提前暴露出来。但是,即使评估者能够自由地去选择所要面对的问题,这类特殊问题也很难被问及。当然,评估者同样能够通过他们的努力对政策性问题的进一步实施产生影响。

资料来源:Eleanor Chelimsky,"The Politics of Program Evaluation,"Society, 1987, 25(1): 26-27.

政策重要性问题

我们已经强调,评估本质上具有实用的和政治的目的。除了我们已经讨论的问题,评估根本上是为了影响决策过程。以下这些问题进一步将评估者的工作与基础研究者的工作区分开来。

政策适当性和政策空间。政策空间(Policy space)是指一套在任何给定的时点都可以获得政治支持的、可供选择的政策集合。在社会项目的设计、实施和评估等方面,可供选择的方案一般都包括在当前的政策空间范围。政策空间随着有影响力的人物为争取其他有影响力人物和普通大众的支持所进行的努力而发生变化。20 世纪末,有关犯罪控制方面的政策空间是由对特定类型犯罪进行长期的、有时是强制实施的社会项目决定的。与此相反,在20 世纪70 年代,该政策的重心是发展以社区为基础的监测中心替代监狱制度,其背景是监狱成为犯罪滋生地,罪犯可以在正常人的世界中得到最好的帮助。

"刑满释放人员过渡期帮助(TARP)"实验证明了政策空间的易变性。在前面章节我们讨论过这个案例。这项实验是在20 世纪70 年代后期实施,目的是评估给释放犯人提供短期经济援助以减少犯罪项目的有效性。无论在佐治亚州和得克萨斯州进行的 TARP 实验本身好坏如何,在评估结果出来之前,联邦的政策空间已经发生了如此彻底的改变,以至于根据实验制定的政策根本无法为人们所考虑。因此,评估者不仅要关注调查项目开始时的现存政策空间,也要密切注意评估实施中社会和政治背景变化对政策空间改变的影响。

时常,人们要尝试一些新项目,尽管政策决策者不一定充分理解项目的政策意涵,却不得不批准实施项目。因此,即使对一些受质疑项目的评估是无瑕疵的,结果仍有可能被证明是不相关的。在新泽西—宾夕法尼亚州收入维持实验中,实验设计者的主要研究问题是:收入维持计划对找工作的消极作用有多大?在实验未完成之前,美国国会制订了不同的收入维持计划,因此,关键问题不再是对找工作的消极作用。相反地,国会议员更关心的是有几种不同的福利制度形式可以综合在一起,由此不再忽视穷人的关键需求,也不再产生出不公正(Rossi and Lyall, 1976)。政策空间和关注点已经发生了明显改变。

因为评估的主要目的是帮助决策者形成和采纳一些社会政策,调查者必须考虑所涉及的不同政策问题和政策空间的限制。一项计划的目标必须与决策者对所关心问题经商议得出的结论类似。如果我们设计了一个严密的随机实验证明税率递减的税收制度可以导致个体生产率提高,而如果决策者关心的是企业家激励和如何吸引潜在投资等问题,那么我们的调查就是不切题的。

因此,要设计合理的评估必须尽可能接触相关决策者,探知他们对待检测项目的兴趣所在。一个有职业敏感性的评估者需要知道当前和未来政策空间所允许考虑的事情。一项革新计划现在不为决策者讨论,但仍会受到检测,因为其可能会成为未来讨论的主题。对计划有效性进行评估的评估者和主办者必须依赖于他们有知识依据的推测去估计将要出现的政策问题,也就是说去设想在政策空间中可能会发生什么可预期的变化。在另外一些计划中,获取决策者观点的过程十分简单。评估者可以参考审议团体的会议(比如政府听证会或立法机关辩论会),可以访问决策者的助手,或者直接咨询决策者。

政策意义。评估是按照社会调查规则进行的,这使其优于其他评判社会项目的方法。但是,评估并不必然提供更多的信息,除非强调从事政策制定、项目计划和管理的人的重要性,也就是说,评估必须提供政策意义。这样看来,评估研究的弱点集中表现在如何确定调查问题和如何解释评估结果(Datta,1980)。这一问题已经超出了纯粹的方法论讨论,解释评估结果包括方法论之外必须考虑的更多事情。要将评估结果效用最大化,评估者就必须关注下面两个层次的政策问题。

第一,比起不重要的问题,受批评的项目需要更好(或说更严密)的评估。像统计显著性水平和效果值设置这样的技术问题应该受到政策和项目因素的制约。这一直是影响评估判断力和敏感度的重要问题,甚至在进行正式有效分析时,这样的问题仍然存在。例如,政策决策者和主办者的考虑因素决定了应当使用个体本位还是项目本位抑或是社区本位的立场。

第二,对于评估结果的评价取决于其具有多大的普遍性,结果对政策和项目是否重要以及项目是否迎合需要(这表现为决策过程中涉及的众多因素)。一项评估结果可能被人们公认具有统计显著性和推断性,但对政策、规划和管理行为却没有实际意义(Lipsey, 1990; Sechrest and Yeaton, 1982)。关于这个内容,我们已经在第10章的"实际意义"中讨论过了。

基础科学模型与政策导向模型。社会科学者经常不能领会阐明模型的目的到底在于改变某一现象,还是发展模型去解释社会现象的差别。举例而言,很多青少年犯罪行为可以在他们社会关系中的男性群体(如父亲、哥哥、其他男性亲属、朋友、邻居、同学等)那里找到原因。这个发现对于犯罪比率的地理学和人种学研究大有贡献。然而,从改变犯罪行为比率角度来说,这个结果并没有用,因为很难设想政府会把改变青年人的社会关系作为一项可以接受的公共政策。除了完全改变青年男子生活环境,把他们带入另一个环境之外,想不到任何其他办法可以改变他们的社会关系。社会政策空间不可能包括为这种目的而重新进行

人口配置的尝试。

相反,对犯罪行为不甚重要的影响因素反而更容易导向一项公共政策,用以改变控制犯罪行为的费用。譬如说,潜在违法者主观上对犯罪后被抓、被判罪、被关押可能性看法的改变可以成为犯罪控制项目的实践性基础。犯罪动机与个体主观可能性的关系很微弱。个体越是相信如果他们犯罪后会被抓、被判罪、被监禁,犯罪行为发生的可能性就越低。因此,如果警察努力监控犯罪行为,法庭制定严厉的审判制度,那么犯罪行为发生的可能性就会在一定范围内得到减少。虽然不具很强的相关性,但这些结果比起前面讨论的社会关系来说,对制定控制犯罪公共政策的实际意义更大。市长和警局长官可以实施提高罪犯逮捕率的计划,起诉人可以尽力监控犯罪行为,法官可以拒绝任何形式的说情。另外,这些政策改变可以传播到潜在罪犯那里,也可以对犯罪率减少发生一点作用。综上所述,可见:基础社会科学模型经常会忽略政策相关性,但是,对于评估者而言,这必须是一个核心关注点。

工程传统的丧失。我们对研究的政策相关性和政策意义的讨论指向一个更为一般性的内容:评估者(实际上是全部应用研究者)和项目方必须长期发展一种"工程传统",这一传统目前在大多数社会科学中已经丧失。工程师们的特点是:他们把"纯科学(pure science)"与对如何使用科学知识解决现实生活问题的兴趣相结合。知道气体受热膨胀、每种气体具有各自的膨胀率是一回事,能够运用这个规律大规模生产高效的蒸气机又是另外一回事。

类似的工程学问题在社会科学发展中也存在。例如,在20世纪80年代,人们对一些既有福利项目十分不满,政策决策者和其他项目方遇到了很多实际问题,但是大家却在如何改进上不知所措。在美国国会中,议员们不断强化的一个共识是:家有小儿帮助计划(AFDC)支付的福利金引发了依赖动机,滋长了福利依赖,阻碍了AFDC计划中的受助者脱离帮助计划、谋求自主就业。在经济学和心理学研究中,有一个著名的理论指向,就是人的动机变化经常会改变行为。那么,"工程"问题就是如何最有效地改变福利项目的激励措施,以此来鼓励受助者减少依赖、寻求职业机会。相应地,健康与人类服务部鼓励政府修改AFDC计划的规则,使受助者有寻求职业的动机。多种涉及动机的建议在随机实验中得到了检验(Gueronand Pauly,1991)。实验检验说明,应该要求在项目中获得福利援助的大人接受求职训练,允许在不减少福利金的同时保留一定份额的工作收入,帮助他们求职。20世纪80年代时,假如工程学传统已经在社会科学中得到了很好地发展与应用,那么,福利改革项目就可以在20世纪90年代末之前的相当长一段时间内得到更好地实施。

尽管我们相信:社会科学的工程学传统是很有价值的一个领域,值得进一步去开发。但是显然,我们的社会科学还不足以完善地建立这一传统,我们的准备十分不足。虽然现有的知识基础在不断增长,但是,如果使工程学传统得到足够的扩展,可能仍然需要几十年的时间。另外,我们并不确切地知道该如何训练这种社会科学"工程师",我们怀疑已有的各种训练模式。也许,必须等到有足够好

的范例出现以供参考之后,这种训练模式才能成型。

我们希望的是,上述对在项目和社会政策真实背景下评估行为发展史的观测报告可以引起评估者对“观察”形势的重视,当他们在从事一项评估并注意在评估过程中发生社会生态学变化的时候。对一项成功的评估行为而言,这种努力至少与采用适当的技术程序一样重要。

对评估本身的评估

随着评估活动变得日益复杂化,评判某项评估是否妥善完成、结果是否得到正确解释变得越来越困难。尤其是对外行和公共部门官员来讲,评价评估的可信度已经超出了他们的能力范围。此外,对于同一个项目的几个评估经常会得出相反的调查结果:怎样使对立的评估论断达成一致甚至成为评估专家的棘手问题。为了使评估结果得到验证,使不同结果之间有足够的交流,人们对下面所说的几种方法进行了尝试。

大规模评估的投资合同和批准常常要求成立由评估专家和政策分析人士组成的顾问委员会来检测评估行为,并为评估者、投资人提供专家意见。设立顾问委员会可以被视为提高评估质量、同时使评估结果更具合法性的一种途径。

还有对评估进行的深入审查,包括对评估数据的再分析。如国家科学委员会有时会组织人员检查评估项目,并综合有关政策利益或重大争论话题的结果。例如,科伊尔,伯勒奇和特纳(Coyle, Boruch, and Turner, 1991)对关于艾滋病教育项目的评估结果进行了检查,并提出了提高类似工作质量的改进措施。类似的批评性评论在其他操作性评估领域中已有所展现,譬如莫斯提勒和伯卢奇(Mosteller and Boruch, 2002)就曾主编过一系列针对教育项目评估的随机实验方法文献。

上面提到的检查工作一般要耗时数年,因此不能满足决策者及时获取信息的需要。要获得更及时的有关评估的评论,就必须进行更迅速的检查和评价。1997年史密斯—理查德森基金会批准资助了一项有希望实现评估及时化的尝试。马里兰大学公共事务学院受委托召集一个由著名评估者和政策分析家组成的名为“蓝带”的调查团,审查和评论对1996年“个体责任与工作机会一致行动”带来的公共福利制度改革的评估过程。福利改革研究评审委员会定期向决策层提交报告,对评估的适当性进行评价,总结它们对于政策的意义。评估审查一般安排在评估报告发布后几个月内进行(Besharov, Germanis and Rossi,1998)。有望实现的是,评审福利改革研究的委员会将会通过呈交给政策决策者的定期报告,来对已有评估研究的完整性进行估量,并指明对社会政策的意义。

最近刚刚成立的国际坎贝尔协会(Internaional Campbell Collabortion)正在实施一项更有野心的工作。这是由关注评估研究的社会科学家组成的一个国际性协会,国际坎贝尔协会已经出版了几期评论,例如关于名为“直接恐惧(Scared straight)”的青少年犯罪预防项目的评论。他们计划定期发表其他评估领域的一系列评论(www. campellcollaboration. org)。

虽然有这样的事例,我们仍然相信典型的项目评估并不简单地受评估领域同行判断的左右。一些决策者可能有权力评判评估的恰当性,但大多数情况下,必须依赖于评估报告的说服力。因此,如后文所要讨论的,评估标准在评估者专业协会中具有重要作用。

评估职业

我们没有自我界定"评估者"的全部名单,也无法给出他们的背景或行动范围的详细描述。在美国至少有大约50 000 ~ 75 000 人在从事全职或兼职的评估行动。我们是把从事社会项目发展和执行的联邦、政府、县和市政府组织的数字与经常从事一种或几种类型评估活动的学校、医院、精神医院和高等院校的数字相加起来,得到了这个估计值。不过,我们不知道有多少人在这些团体中做评估工作,我们也不能预测有多少大学教授和其他人在与营利性和非营利性公司合作做评估研究。实际上,全职或兼职评估者的确切数字或许是我们预测的这个最小值的2 ~3 倍。

显然,评估者在差异巨大的社会项目领域中工作,他们花在评估活动中的工作时间也各不相同。对评估者角色的界定是模糊不定的。

一方面,人们将评估当作一项附属性行为去完成。有时候仅仅是为了符合立法的或制定规章的需要而进行评估活动,许多地方学校的制度管理显然属于这种情形。按照国家或联邦提供资金资助的要求,学校必须任命一名"评估员",他就以这种身份处理教学或管理事务。这个"评估员"往往既未受过专门训练也没有什么经验,不具备胜任这一职位的资格。实际上,有时候这个"评估员"既不是一个好的教师,也不是一个德高望重的管理者,让他承担评估任务只是一个安全保守的决策。另一方面,在大学的评估研究所和社会科学系、在实用社会调查公司的私人非营利部门,有很多受过训练并具有多年实践经验的全职评估专家在评估领域的一线工作。

实际上,评估者或评估研究者这个一般性的标签掩盖了该领域的异质性、多样性和无组织性。评估者们没有从业许可证或执照,所以一个人作为评估者的身份不能保证他和另外一个同样被称作评估者的人分享核心知识。评估者之间很少相互影响、沟通,尤其是跨项目的交流更少。该领域中最主要的综合性组织——美国评估学会仅有几千名成员,而订户最多的交叉学科刊物——《评估评论》也仅有几千读者。在项目评估中,评估者之间的社会关系网络同样作用微弱,大多数人与全国的或当地的评估专业组织没有联系。

总的说来,按照社会学家对这类团体特征描述的标准来说,评估至少尚未成为专业,更近似于"类团体",由一大群未经正式组织起来的人组成,成员资格经常变动,成员之间在承担任务的范围、个体能力、工作场所及观点等方面差异很大。评估的这个特点引发了我们下面的讨论。

评估从业人员的多元化及其后果

评估研究有着丰富而多元的学科支持，所有社会科学的准则（如经济学，心理学，社会学，政治学和人类学）对评估领域的发展都有帮助。接受这些准则训练的人对评估研究的概念基础和方法论形成都有贡献。人类服务的专业领域也对评估发展做出了贡献：人们在多个与社会科学密切相关的人类服务专业（医学、公共健康、社会福利、城市规划、公共管理、教育学等）方面的训练促进了方法论的发展，并开展一些重要的评估活动。此外，统计学、生物统计学、计量经济学和心理测量学等实用数学也发展了评估的测量和分析的重要思想。

学科之间大幅度地互相借鉴。举例而言，虽然经济学在传统上不是一个以实验为基础的社会科学，但经济学家在过去几十年内已经设计和进行了相当多的大规模随机实地实验，包括职业培训、收入维持、住房津贴和国家健康保险等实验。社会学家和心理学家则从计量经济学中借鉴了很多内容，特别是时间序列分析方法和联立方程模型的运用。社会学家的贡献则是督导组织绩效中使用的许多概念和数据搜集的程序。心理学家的贡献是对时间序列分析的非连续性回归分析。心理测量学家提供了适用于所有领域的测量理论的基本思想。人类学家提供了在定性的实地工作中使用的基本方法。评估学的词汇表实际上是这些学科的一个混合物。本书后面所附的参考书目是评估领域具有多学科特征的一个证明。

从理论上来说，这个领域发端的多样性是其魅力所在。然而，从实践上说，评估者即便要维持（更不用说拓宽了）他们的知识基础，也需要成为通才的社会学家和终身的求知者。进一步说，这个领域的多样性也使评估者可能出现“不适当”的方法选择，他们有时因此受到批判。显然，每一位评估者不可能成为社会科学所有领域的专家，不可能成为精通每一项方法论程序的专家。

现在，对具有宽广知识基础和多才多艺“通才”评估者的需要仍没有改变。这种情况意味着：评估者因为他们知识基础的局限必须时常放弃一些工作机会，他们必须使用“近乎完美”的方法而不是不熟悉的但却是最合适的方法去工作，评估主办者和评估事务管理者必须在决定任务的承接者和制订工作安排的时候精挑细选。这也就是说，评估者有时候需要频繁使用顾问，并征求同行们的意见。

在评估行业中可以提供一系列的学习机会，使人们保持与技术发展同步，扩展个体能力范围，如区域性和全国性会议，以及由专业协会开设的教育课程。目前，在数千名评估从业者中，仅有一部分人参加了专业评估协会，并受益于学会提供的机会。

评估者的教育

评估者所接受的教育方式差异很大，这加剧了评估研究领域的分散化特征。评估领域中几乎没有人可以从一个评估机构的底层做起，最终获得重要的职务

和相应报酬。大多数评估者在社会科学系或专业学院中接受过正规的研究生培训。评估的多学科性决定了在某个单一学科内全面参与评估活动方面的训练是不够用的。在少数大学里,很多系已经有了包括研究生课程讲授的跨学科计划。在这种计划中,研究生们要学习心理系的实验建构和测量课程、经济系的计量经济学、社会学系的调查设计与分析、政治学系的政策分析等内容。

然而,跨学科训练计划既不常见也不稳定。在典型的研究生训练和研究导向的大学中,传统学科是有权力的单位。跨学科计划中的不同学科教职员的联盟很容易半途而废,原因在于那些学科通常对这样的尝试态度冷淡,以至最后,教职员们又回到了他们各自原来所属的学科。其结果就是:尽管很需要多学科综合评估研究训练,但这种训练常常本质上仍然是单一化的。

甚至在学术性学科内部,实用工作也常常不如纯研究、基础研究那样受到重视。从而,造成与评估相关的能力训练经常受到限制。心理系可以开设很好的实验设计课程,但并不能像对待实验室研究那样充分考虑实地实验进行中的特殊问题;社会学系可以教授社会调查研究课程,但从不涉及社会项目中特殊群体的数据搜集问题。同时,实用工作在研究生院的较低地位也使评估无法成为学术论文或学位论文的研究主题。我们所能给的意见就是对评估职业有兴趣的学生必须坚持己见。这类学生通常必须自己制订个体学习计划,去学习不同系开设的课程;必须坚持用实用性课题做自己的学术论文或学位论文;必须抓住大学研究生院以及学术圈中一切机会,参加补充知识和训练技能的正规课程。

另一种训练路径是职业学校。职业学校可以培训评估者使其获得在该领域内的地位,诸如此类的像公共健康与医疗照顾学校的项目产生了从事健康服务评估的人员。实际上,除了 MBA 项目之外,在一定时期内这些职业学校成为许多评估者的受训地点。

这些培训项目也各有局限。对他们的一种批评意见是——他们在观点上过于商业化,这使得有些培训项目不能提供概念深度与广度拓展,以及让学生们在社会项目领域中游刃有余并掌握技术革新能力。此外,特别是在硕士研究生阶段,许多职业学校需要有大量的必修课,因为他们的地位和资金来源依赖于专业团体的认证,而这些专业团体所认可的都是一些培养社会福利硕士(MSW)、公共健康硕士(MPH)、工商管理硕士(MBA)人才的一般性训练。许多培训项目因此给选修课留下了极少的时间,在技术训练方面的课程实在有限。在职业学校中,对评估者的训练逐渐从硕士课程转移到博士课程。

同样,在许多大学,社会科学的传统学科视职业学校的教职员和学生为二等公民。这种精英主义使职业学校的学生被孤立。他们不能在所属的社会科学研究学院中同时学习几个社会科学学科的课程并接受相关培训。在职业学校中受训的学生,在硕士研究生阶段经常丧失对在某一特定项目领域主要知识进行集中技能训练和获得职业执照的机会。明确的补救办法是进行更深入的研究工作或者在从事评估职业时抓住机会学习其他技术、技能。

我们没有比较训练路线之间的优劣,每一种路线都有利弊。看起来职业学

校正逐渐成为(至少是部分成为)培养评估者的主要途径,原因是研究生社会科学系科都不愿推进专门的应用研究项目。但职业学校之间在教授内容上差别很大,尤其在所强调的评估方法上各不相同,因此造成了这个领域多样化现象继续存在。

来历多样性的影响

培养评估者的多种教育途径造成了这个领域凝聚力的匮乏。这造成了(至少是部分影响了)对评估的定义分歧,以及评估社会项目时方法选择的差异。当然,也有其他因素影响着这种差异。

这种差异一部分与评估者是在职业学校还是在社会科学系接受教育有关。譬如,比起接受社会学训练出身的评估者来说,从职业学校出来的评估者更可能把自己作为项目工作人员的一部分,并认为应该优先考虑协助项目经理完成任务。因此,他们可能强调可以改善项目日常操作的"塑造性评估"。

这种差异也与社会科学系科之间的差异、职业学校之间的差异有关。接受政治科学训练的评估者经常遵循政策分析导向,为立法者和高层行政官员尤其是政府管理者提供帮助。人类学家如人们所料,倾向于定性方法,经常关注评估结果中的目标人群的利益。对于心理学家,他们的学科准则重点在于小范围实验,与此一致,他们关心评估中因果推理的有效性更甚于项目实践的一般化。相反,社会学家经常更关注一般化的可能性,为达到这一目的他们更愿意降低因果推论的前提条件。经济学家则可能用不同的方式工作,他们依靠微观经济学理论来指导他们的评估设计。

在不同职业学校受教育的人之间也存在这样的差异。在对儿童早期教育项目的评测中,教育学校出身的评估者重视的是教育能力测试,而社会工作毕业生重视的是对于儿童情绪状况的评价和家长对儿童行为的报告。公共健康学校出来的人最感兴趣的是预防性实践、医生出诊的频率和医疗保险的期限等的医疗照顾管理项目。

我们很容易夸大每一种设计和完成评估的指导准则及专业的特殊见解,这种偏好取向也不乏例外。实际上,评估爱好者们最喜爱的一个游戏就是从一篇文章的上下文来猜测研究者的学科背景。尽管如此,学科的和专业的差异仍在评估领域中造成了很大程度的对立冲突。从认识论到方法和主要评估目标的选择,评估者都持有各自不同的观点。下面将简要介绍一些主要分歧。

指向主要项目方的取向。正如本章前面所提到的,一些评估者相信应该以帮助项目管理者改善项目为方向。这种观点主要将评估的目的看作是为项目管理者提供咨询,这就使技术性帮助和评估之间的差异变得模糊。根据这种观点,项目改善程度是评估是否成功的标准。这种评估倾向试图避免对项目价值的判定,使大多数项目方都能够在评估的帮助下开展工作(参见 Patton, 1997,他引发了有关这个问题的持续性讨论)。

另外一些人则坚持评估的目的是帮助项目受益者(项目的目标群体)增权。

这种评估观点认为,让项目的目标群体联合起来解释项目并对其进行评估可以使目标群体增强自主性,提高他们对自身效力的评价(Fetterman, Kaftarian and Wandersman, 1996,其中有对于这种方法的例证)。

另一种极端的观点则认为评估者应该主要为评估提供资金的项目方服务。这种评估体现投资者的想法,采纳他们有关项目目标和项目结果的界定。

我们的观点已经在本章前面表述过了。我们相信评估应该考虑主要项目方的看法。通常,这种评估所具有的契约性要求将主要注意力放在评估主办方对项目目标和结果的界定上。然而,这样的要求并不排斥其他观点。我们相信详细论述评估中产生的观点以及明确指出评估中所包含的主要观点,都是评估者的义务。如果一个评估具备接纳多种观点的条件,那么就应该使用多元化的观点。

认识论差异。"文化战争"已经在人文和社会科学中展开,也影响到了评估研究。后现代的知识论认为,实证主义认识论已经被知识相对论所替代,反映在评估领域就是——社会问题是社会结构的问题,知识不是绝对的,存在着不同的"真理",每一种知识对于一定的视角而言都是正确的。后现代主义者偏好定性研究方法,可以获得丰富的"自然"数据,在评估上则偏爱项目的全体人员和目标人群(参见 Guba and Lincoln, 1989,后现代评估的最重要代表)。

认识论的后现代立场并于知识本质的理念是不同的。但是,仍有一些观点坚持认为真理不完全是相对的。譬如说,大多数人相信贫穷是一个社会结构的概念。但同时也有人确信年收入的分配可以通过大多数社会科学家都赞同的研究操作来描述。也就是说,在某一收入水平上是否会被称作"贫穷"是有关社会公正的问题,完全可以在已知样本误差的条件下估测某个收入水平的家庭户数。就是说,研究者在经验调查结果上的分歧主要是方法或测量误差,并不包含不同真理体系的差异。

我们在本书论述的观点显然不是后现代的。我们相信方法与评估问题是严格匹配的。对一个给定的研究问题,有好方法也有坏方法。实际上本书提供的主要信息是:只有选择最好的方法进行问题研究,才可能获得最可信的结果。

定性与定量的区别。偏好定性研究方法与偏好定量研究方法是评估研究圈子内同时存在的另一差异。有一部平淡乏味的著作就是围绕这一问题展开的。一方面,定性方法的拥护者强调我们需要个性知识,需要熟知一个项目在获得有关项目影响的有效知识方面的具体表示。定性评估者倾向于走向塑造性评估,也即是说,通过为项目管理者提供有关项目的信息来改进项目工作。与此相反,定量方法取向的评估者经常认为这个领域主要关注的是总和性评估,其焦点是发展对项目特征、进程和影响的测量,使项目的有效性得到高可信度的评估。

争辩常常掩盖问题的关键所在,每种方法都有其效用,方法的选择往往由手头的评估问题来决定。在本书中,我们试图确定每一种观点的适用性。正如我们所强调,定性方法在项目设计中起决定性作用,是一种重要的检测项目的手段。相反,定量方法更适用于净效果的评估以及社会项目产出功效的评估(对于

定性与定量方法的对比讨论,参见 Reichardt and Rallis,1994)。

因此,在没有具体确定评估问题的情况下,讨论哪一种方法更合适是无意义的。使方法适合研究目的才具有决定意义。使一种方法抽象地与另一种方法对立,结果只能将这个领域毫无意义地一分为二。即使某种方法最热情的拥护者也会承认,每一种方法对社会项目评估都具有贡献(Cronbach, 1982, Patton,1997)。实际上,如果通过不同方法得到的结果是一致的话,被称为"三角测量"的多种方法的使用,可以增强评估结果的有效性。使用多种方法是消减各种偏见和测量误差的一种方法(对这一观点的进一步讨论,参见 Greene and Caracelli, 1997)。

正如我们所见,这个问题不仅是哲学上的,也是战略上的。评估主要对政策和项目的塑造及修改发生作用,正像我们强调的那样,是具有强烈政治特征的活动。正如切里姆斯基(Chelimsky, 1987,p. 27)所指明的那样,"仅凭一个案例研究就参与到激烈的政治争论中去,是非常不明智的"。

工作安排的多样性

评估领域的多样性还体现在评估活动所面对的环境和评估者所在的官僚体系的多样性。首先,针对评估工作安排有两种对立观点,或称之为"局内人与局外人"的争论。一方观点认为,评估者最好尽可能远离项目管理者及其职员的影响,以使他们的地位更安全、独立。另一方认为,与政策和项目职员保持长期接触可以增强评估者工作效果,使他们更好理解组织目标和活动,培养相互之间的信任,从而提高评估者的影响力。

其次,无论评估者是局内人还是局外人,假定评估者与公司或者学术界应该有怎样的关系本身是一个问题,他面对项目职员和项目各方的角色也可能是含混的。但这只是衡量评估者建构合适工作关系挑战的尺度之一。

第三是对部分评估者所属组织地位(standing)的关注。和大学一样,评估者的工作环境会被等级化,并以许多标准来评定,只有很少几个大的评估组织构成了公认的顶尖集合。是做大池里的小鱼好,还是相反,同样是评估领域的一个问题了。

下面的讨论更多是研究者的观点,而不是经验性的研究发现。我们的观点可能是错的,但无论评估者受到多大的争议,争议本身就是一件平常事情。

内部和外部评估

过去,一些有经验的评估者试图证明评估不应该在对项目负责的组织内进行,而应该由局外团体来完成。"局外者"评估的一个理由就是局内人和局外人在训练及假设能力层次上的差异。这种差异已经被缩小了。评估研究者的经历一般都对应着三种形式之一。直到20世纪60年代,很大一部分评估研究都由隶属于大学的研究者或研究公司完成的。从60年代晚期开始,不同项目领域的公共服务机构雇佣研究者去做更多的内部评估(in-house evaluations)。同时,由营利性团体完成的评估的数量也在大幅度增加。由于这两种组织中研究职位的增

加以及理论性工作市场的缩小,越来越多的受过良好社会学和行为学训练的人投身于公共机构和营利性公司的研究工作。

现在的证据还远远不能弄清内部评估或外部评估哪一个具有更高的技术质量。但技术质量不是唯一的标准——实用性也同样重要。一个在荷兰进行的有关内部和外部评估的研究显示:内部评估可能对组织决策的影响更大。正如凡德沃和伯拉斯(van de Vall and Bolas, 1981)所说的那样,随着内部研究发现利用率的提高,这一部分研究人员在影响社会政策方面的重要性加强。这种状况可能部分存在于内部研究者和政策制定者之间交流的加强以及认识与帮助效力标准的平衡上:"在运行期间,这意味着社会政策的研究者应该在完善方法所花费的时间以及将结果转化为政策尺度之间寻找一个平衡"。他们的数据表明当前的"局内"社会研究者要比外部调查者在实现有帮助性的目标方面处于更有利的位置。

随着职员能力的增加以及评估透明度和详细度的提高,现在已经没有理由将一个组织化的安排完全凌驾于另一个之上。然而,当一些工作的时机被误导或在评估中经常性忽略评估者地位的时候,仍然会出现许多批判观点。重要的是,任何评估的目的都会在技术性和实用性之间寻求一个平衡点,而这些目的对内部评估和外部评估来说可能是不同的。

组织的角色

不管评估者是局内的还是局外的,他们都需要对项目主办者与项目职员的角色有清楚的认识。评估者对他们角色与责任的完全理解是成功完成一项评估最主要的支持因素。

同时,这个领域发展的不均衡性导致很难总结出发展和维持适当工作关系的最佳途径。一个常用的技巧,就是适当利用咨询小组或一个乃至更多的顾问来督导评估并为研究发现提供可靠的氛围。这种咨询小组或顾问的工作方式依赖于内部或外部调查的采用,以及评估者和项目职员的经验程度。例如:由联邦机构和大基金会发起的大型评估往往配备能够评定工作质量、数量和方向性的咨询小组。一些带有小型评估部门的公共和私人健康福利组织具有可以向评估者提供技术意见以及向机构领导者提供评估部门活动可行性建议的顾问。

有时候,咨询小组和顾问仅仅是个装饰,如果只有这个作用的话,是不值得提倡的。不过,当成员都积极投入时,咨询小组就能在各领域间评估联系方面、解决项目和评估职员之间的争端方法以及评估发现利益威胁群体对评估者进行攻击时保护评估结果方面,起到特别的作用。

"精英"组织的领导角色

一小部分评估者,数量估计不超过 1 000 人,因为他们参加评估的规模和他们所供职组织的规模组成了该领域的"精英"。他们在某些方面同在重要医学院医院里工作的杰出医生非常类似。他们的人数和评估量在数量上都不大,但在

为该领域建立标准方面却非常权威。他们工作的方式以及他们所在组织的行为标准都是职业特性的重要说明,被业界的其他人视为典范。

这种具有强大的技术力量、能进行全国性的或是大规模评估的组织非常少。但在受瞩目程度和评估花费方面,这些组织占据了战略性的位置。绝大多数规模较大且持续几年的联合评估都分配给了一小部分的营利性研究组织(诸如阿卜特联盟,计量政策研究,以及 Westat,这里只是列出一小部分)和非营利性研究组织和学校(例如贝特勒纪念学院(Battelle Memorial),兰德公司,研究三角机构,城市研究所,以及人力发展研究公司)。少数含有研究所的大学——芝加哥大学的国家民意研究中心(NORC),威斯康星大学的贫困问题研究所,以及密歇根大学的社会研究所——也接受一些大规模评估工作的委托。另外,为评估研究提供资金的联邦机构评估部门以及一些大的全国性基金会,在他们的职员中都有大批受过严格训练的评估者。联邦政府可能是有经验评估者最集中的地方,直到最近,仍然有大批评估专家被调入美国审计总署的项目评估与方法局去处理"看守者"组织的事务,涉及从审计到评定项目的落实以及估计联邦优先权的影响。

这些盈利和非营利性精英组织的一个重要特征就是对工作质量给予从不间断的关注。这种情况部分来自于他们早期的努力受到过严厉的批评。起初,他们的工作在技术性上比不上学术机构(Bernstein and Freeman, 1975)。但当他们开始统治这个领域并开始进行大规模评估活动时,他们发现评估主办方在选择委托对象时越来越注重技术能力,于是他们的努力立刻在方法论上发生了显著的转移。在具备了有能力的职员后,他们仍在努力争夺在工作中按受过最佳训练的人。同时,出于自身利益的需要,他们鼓励职员在行业性杂志上发表文章、积极参加行业组织,并参与到改善评估技术的前沿研究之中。在向职业化不断迈进的过程中,这些组织自然就成为了领导者。

评估标准、准则和伦理

即便评估领域不能在通常意义上被定义为一个有组织的职业体系,但的确正逐步地走向专业化。由于评估领域越来越专业化,许多评估者开始向他们所属的职业联合会施加压力以求能在评估工作中或在与评估投资方和其他主要项目方的谈判过程中将起指导作用的标准加以规范化并公开。例如,如果他们在为争取自由出版评估发现的权利而引发的争论中,引用一套已被出版的有关自由出版权的实践标准,那将是非常有利的。另外,几乎每个实践中的评估者都会遇到需要道德评判的情况。例如,在评估者对防止虐待儿童的项目进行调查时,他有义务将调查采访过程中发现的某个家庭虐待儿童的现象进行具体披露和报告吗?已出版的标准或实践指导也能给那些宣称提供与之相符评估服务的评估者提供合法性。

在为评估者提供实践指导方面已有两个大的成果。在美国国家标准机构(ANSI)的支持下,教育评估标准委员会(1994)出版了《项目评估标准》一书,现在已经是第二版了。委员会主要由几家职业协会的代表组成,包括美国评估协会、美国心理学协会以及美国教育研究协会。最初,这个委员会只是为了处理教育性项目而成立,现在已经渐渐将范围扩展到所有项目评估领域。该标准包含了非常广阔的主题,从评估合同的准备到处理人权的条例,再到分析定量及定性数据的标准等。每一个标准都以具体案例加以说明,以指导如何在特殊的个案中运用这些标准。

第二个努力就是被美国评估协会采纳的《评估者守则》(*American Evaluation Association*, Task Force on Guiding Principles for Evaluators,1995)。守则总结了5条指引评估者实践工作的原则。具体内容在专栏12—E中列出。

- 系统调查:评估者对评估对象进行有计划的、以数据为基础的调查。
- 能力:评估者向项目方提供胜任的表现。
- 正直/诚实:评估者保证在整个评估过程中的正直和诚实。
- 尊重他人:评估者尊重接受调查者、项目参加人员、客户以及其他项目方的安全、尊严和自我价值。
- 对大众和公共福利的责任:评估者明了并重视与公共福利有关的利益和价值的多样性。

这五条准则虽然没有达到联合委员会工作所要求的详细程度,但仍然在《评估者守则》一书中进行了详细的设计和讨论。这样的概括性原则能起到多大作用仍就是一个疑问。一个遇到特殊道德问题的评估者仍然在这些原则中很难找到指引(参见 Shadish, Newman, et al., 1995)。

我们期望发展一套能够给评估者提供针对性建议的实践标准和道德准则,但这仍需要一个长期的过程。评估体系的多样性使吸收接纳一套标准变得非常困难,因为任何以某标准来指导的实践都可能与其他组织认为正确的实践发生抵触。标准的发展将会因为"个案规律"的存在而变得非常超前,所谓"个案规律"就是在处理过程中引用了某些准则的特殊案例的积累。然而,无论教育评估标准委员会的标准,还是美国评估协会的指导原则都不是强制性的,只是个案规律发展的一个寻常过程。

在这样的评估标准和伦理准则建立以前,评估者只能依赖于当前被业界认可的总体原则。关于项目评估道德标准的许多有益讨论可以在纽曼和布朗(Newman and Brown,1996)的相关著作中找到。

评估者必须理解:评估者守则并不能替代大多数社会服务机构和大学所认可的伦理标准。大多数社会研究中心和几乎所有大学都有常设的委员会来处理关于人类社会服务的相关研究立项或申请,大都要求研究方案先提交给他们进行审查。几乎所有这类审查评论都关注被研究者的"知情同意",这一准则要求研究者让被研究对象在参与研究之前,有充分的知情权,应该清楚地了解参与研究所可能带来的风险,这是研究操作的必要步骤之一。另外,大部分职业协会

(例如美国社会学会、美国心理学会)也有自己的一套伦理规范,能够为一些职业问题提供有价值的指导,例如适当地感谢合作对象,避免剥削研究助手等。

如何将这些准则应用于评估之中,难易参半。如果是那些大家在各种情景下普遍遵守的伦理标准,就易于被应用于评估实践;相反,如果是那些可能与研究需要相冲突的伦理规范,则难以执行。例如,一名急需业务的评估者很可能会投标一些必须应用自己不熟悉的方法才能完成的评估项目,这一行为明显与评估准则相违背。另外,评估者也会担心他准备使用的评估方法是否能够提供足够的相关信息,以使参与者明白参与项目的风险。在这类情况下,我们的建议是:评估者应该主动咨询其他有经验的评估人员;同时,在任何情况下,争取避免采取与评估规则相冲突的行动。

专栏 12—E 美国评估协会的指导原则

A. 系统调查:评估者对评估对象进行有计划的、以数据为基础的调查。

1. 评估者在工作中应坚持遵守最高的、适用的技术标准,不管这项工作在本质上是定量的还是定性的,这样才能提高评估信息的准确度和可信度。

2. 评估者应该和客户一起探寻评估中访谈问题和方法的优缺点。

3. 在陈述工作情况时,评估者应该准确和详尽地展示方案和具体操作,以便大家了解、评论。应该充分展示评估及其结果的局限性。评估者应该以清晰的、合适的方式讨论对评估结果有重大影响的价值、假设、理论、方案和分析等,从最初的概念化到最终的结论发现。

B. 能力:评估者向项目方展现能够胜任的能力。

1. 评估者应该具有(或者确信评估小组具有)胜任评估任务的学历、能力、技能和经验。

2. 评估者应该在自己职业训练和能力的范围内工作,对超出范围的评估工作应予以拒绝。当无法拒绝委托时,评估者应该清楚地认识到评估中任何重要的限制都是决定性的,评估者应该尽一切努力获得完成任务的能力,或通过专业人员的帮助来获得能力。

3. 评估者应该不断提高自己的工作能力,从而在评估中做得更好。这种不断的职业培训既包括正式的课程、自学、自我工作的评判,也包括同其他评估者共事时学到的技能和专业知识。

C. 正直/诚实:评估者保证在整个评估过程中的诚实和正直。

1. 评估者应该与客户和有关项目方对评估的费用、要完成的任务、评估方法的局限性、可能得到的评估结果以及从评估中得到数据的功用展开诚实的谈判。对这些事的讨论和详细说明首先是评估者应尽的责任和义务。

2. 评估者应该记录一切在最初谈判制订项目计划之后的变化以及变化的原因。如果变化极大地影响了评估结果,评估者(除非有更好的理由,在继续调查以前)应该及时通知客户和其他重要项目方,报告发生的变化和可能导致的影响。

3. 评估者应该判断与评估结果相关的利益,包括自己的、客户的以及其他项目方的(包括财务上、政治上的以及经历上的)。

4. 评估者应该公开所有可能与评估者身份发生冲突的、曾经涉及过的角色和关系。在评估报告中指出必须提到所有相关冲突。

5. 评估者不应该误传任何调查的进程、数据和发现。在合理的条件下,应该努力防止或纠正其他人对于评估工作的错误言论和行为。

6. 如果认为某些步骤或事件引起了对评估信息或结论的误导,评估者有责任将发生的事情及原因向客户通告。如果与客户的讨论无法解决问题,评估者可以在可行的情况下正当合法地拒绝评估。如果不可行,评估者应该咨询同事或有关项目方以求得其他合理的解决办法(可以有自由选择,但不是必须的,可以有更高层的讨论,一封表示异议的封面信或附录,或是拒绝在最终文件上签字)。

7. 除非有不得已的理由,评估者应该公开评估使用的所有资金来源以及评估的委托方。

D. 尊重他人:评估者尊重接受调查者、项目参加人员、客户以及其他项目方的安全、尊严和自我价值。

1. 只要可能,评估者必须遵守职业道德标准,告知参加评估可能产生的风险、伤害和额外的负担,要求达到的共识以及保守秘密的限制和范围。这种标准的实例包括:保护人权的联合准则,或诸如美国人类学协会、美国教育研究学会或美国心理学协会的道德准则。虽然这个准则并不打算扩展适用范围,但评估者还是应该在任何可行条件下或被要求的地方遵守。

2. 因为必须说明评估得到的否定或批评性结论,所以评估有时候会产生一些伤害客户或项目方利益的后果。在不破坏整个评估的正直诚实的前提下,评估者应该努力增大获益并减小不必要的损害。评估者应该谨慎判断从评估中获益的时机以及会造成风险和伤害而应该忽略的环节。只要可能,在评估前的谈判中,对上述问题都要有所预计。

3. 因为评估经常会消极地影响资助者的利益,所以评估者应该使用尊重项目方尊严和自我价值的方法进行评估并交流评估成果。

4. 只要可行,评估者应该努力保证评估的公正性,使为评估付出的人获益。例如评估者应该确定为评估提供数据而承受压力和风险的人是完全自愿的,而且有足够的知识和最大的机会获得评估所带来的好处。在不危及评估公正性的前提下,应该告知评估参加者如何获得他们有权得到的服务,即使他们不参加评估。

5. 评估者有责任区分并尊重参加评估者的不同情况,例如他们在文化、宗教信仰、性别、残疾、年龄、性别取向以及民族等方面的差异,评估者在计划、实施、分析以及汇报评估时,应该留意这些差异潜在的意义。

E. 对大众和公共福利的责任:评估者明了并重视与公共福利有关的利益和价值的多样性。

1. 在计划和汇报评估时,评估者应该考虑有关项目方对被评估事件全方位的重要观点和利益。评估者在忽略有重要价值的意见或重要团体的观点时,应该谨慎地思考其合理性。

2. 评估者应该不仅考虑当前的运作及评估结果,还要考虑更宽泛的假设、联系以及潜在的影响。

3. 信息自由在民主国家是必需的。因此,除非有强制性的理由,评估者应该允许所有相关项目方获知评估的信息并在资源允许的前提下积极向项目方发布信息。如果不同评估结果以适合不同项目方利益的形式传送到他们手中,就必须确认每个项目方都了解其他交流形式的存在。被处理过并传送到他们手中的信息应该始终包括所有重要的、可能对他们的利益有影响的结果。无论如何,评估者应该努力在不影响准确度的前提下使结果尽量简单明了,以使客户和其他项目方理解评估的过程和结果。

4. 评估者应该在客户需求和其他人需求之间维持平衡。评估者有必要同提供资金或要求进行评估的客户维持特殊关系。由于这种关系,只要可行、合适,评估者必须努力满足客户合法的要求。但这种关系在客户利益与其他人利益发生冲突或当客户利益与评估者出于对大局需要、能力、公证性以及对人的尊重等方面的考虑发生冲突时,将会把评估者置于困难处境。在这种情况下,评估者

应该清楚地区分并与客户和有关项目方讨论这种冲突，尽可能地解决冲突；在冲突无法解决的情况下，决定是否继续评估工作以及弄清任何可能对评估产生影响的重要限制因素。

5. 评估者有维护公众利益的义务，这对公众性基金支持的评估尤其重要。评估中，不应该忽略任何对公众利益的威胁，因为公众利益与其他团体的利益是完全不同的。评估者在把社会利益作为整体考虑时，就不得不忽略对项目方利益的分析。

资料来源：American Evaluation Association, "Guiding Principles for Evaluators: A Report from the AEA Task Force on Guiding Principles for Evaluators," by D. Newman, M. A. Scheire, W. Shadish, & C. Wye. Available from http://eval.org/Evaluation Documents/aeaprin6.html. Reprinted with permission.

评估结果的利用

根本上讲，评估的价值一定要用他们的实用性来判定。因此，在评估结果的应用方面要投入大量的思考和研究。首先，我们来讨论利用评估结果的三种通用方式，这对展开分析和实际运用都有帮助(Leviton and Hughes, 1981; Rich, 1977; Weiss, 1988)。

首先，评估者重视对评估结果的直接利用。直接利用就是指能够直接为决策者和其他项目方使用和发挥其特定的功用。例如，健康保持组织的病人在住院时间方面要比在临时医疗中心的病人少的数据，就被国会和健康政策制订者在贫困人口医疗照顾项目中使用(Freeman, Kiecolt, and Allen, 1982)。最近，人力发展研究公司对家有小儿帮助项目中的弃权者进行了工作福利制度研究，他们的工作非常出色，对国家如何改革福利制度产生了影响(Gueron and Pauly, 1991)。

第二，评估者还重视评估的概念性利用价值。正如里奇(Rich, 1977)定义的，概念化利用是指评估从总体上对人的思考所产生的影响。例如，当前控制健康和福利服务费用的努力至少部分是由对服务实效和收益率的评估所引起的。这些评估没有导致具体的项目或政策的采纳，但却证明了当前提供健康服务的方式是昂贵且低效的。

第三，另外，还有一种使用类型是劝导性利用，即列出评估结果的效用，用来支持或反驳某种政治立场。换句话说，是为了保护或攻击现状。例如，里根政府在回击对削减社会项目的指责时最常用的理由就是大多数项目评估中缺乏清晰的发现和确切的影响。劝导性利用和在政治演讲中插入例证类似，不论这种例证是否恰当。大多数情况下，评估的劝导性利用不由项目评估者和项目方所控制，而且以后也不会有什么关系。

可以对评估进行直接利用吗

对评估实用性的失望，表面上来看似乎要归咎于其有限的直接用途。但对直接利用的系统研究也只是最近二十年的事情。这些新近的尝试，直接挑战了

之前认为评估没有直接功效的观点。

例如,在仔细研究了美国教育部发起的评估直接利用后(Leviton and Boruch, 1983),他们发现了大量评估结果导致重要项目变更的实例以及更多的在决策过程中不是唯一影响要素但却的确发生影响的实例。

切里姆斯基(Chelimsky, 1991)也提及了几个社会科学研究中为公众策略的发展提供决定性信息的案例。可惜的是,大规模的评估在已出版著作中占了统治地位。许多小规模评估,特别是诊断性和塑造性且在改进项目方面起过直接作用的评估,却很难找到有关记载。

总之,衡量直接利用评估结果的合适尺度还有待斟酌。很多人,包括评估者和评估的潜在使用者,对此都持很乐观的态度。按照他们的看法,我们应该加大对评估实际利用的幅度。实际上,这种建议刚好与不断增加的对评估结果直接利用的现实相一致。但另一方面,适当地重视评估的概念性利用也非常重要,同样不容忽视。

评估的概念性利用

毫无疑问,每个评估者都曾有过这样的美梦:美好的世界带着赞扬来接受评估发现,并立刻把结果投入使用。大多数这样的梦想都仍然停留于设想阶段。当然,我们可以反驳——评估的概念性利用经常为政策和项目发展提供重要的帮助,不应该等闲视之。概念性效用可能并不能够为同伴或项目方亲眼所见,但评估的这种利用作为整体或重要部分对社会产生深刻影响。

“概念性利用”是指评估对政策、项目和操作步骤发生间接影响的多样途径。既包括使个体和团体对当前的社会问题变得敏感,也包括通过一系列评估的累计结果来影响未来的项目和政策发展。

评估通过记录社会问题的发生、普遍性以及区分社会问题的不同特征而扮演了“光敏处理”的角色。诊断性的评估活动,如第4章描述的,为家庭体系的变化、无业人员的定位与分配的评判性信息以及其他有社会意义的描述等,都提供了越来越清晰和精确的理解。

有影响的评估同样有概念性功效。一个具体的例子就是当前医疗救助政策发展中的“夹缝”团体。为贫困者提供医疗救助项目的评估发现,最贫困、够资格领取诸如国民医疗补助等公共福利的人通常能得到足够的健康服务。而稍好于他们的“夹缝”团体因不够资格领取公共服务,就落入了接受社会服务与自给自足的夹缝里。他们想接受服务,但这无疑是困难的,如果他们患重病,就会成为社区医院的沉重负担,医院既不能把他们拒之门外,也无法从病人或政府得到任何补偿。对于这些贫困者的关注还在增加,因为他们被普遍排斥在身体健康、心理健康以及社会服务项目门外。

一项研究产生长远影响的有趣例子就是现在最经典的、有关教育机会的科尔曼报告(Coleman et al., 1966)。这项研究的发起是在1964年由国会委托教育办公室提供美国少数民族学生接受教育机会质量的信息。这项报告的实际影响

非常深远：改变了教育环境好坏特征的传统观念，将政策和项目的重心从财政支持转移到改善教育计划上（Moynihan, 1991）。

评估结果的概念性功效通常以间接和难以探寻的方式通过多种途径渗透到政策和项目中。例如，科尔曼给教育办公室的报告并没有成为政府公开发行的办公室畅销读物，几乎没有什么人能够一页一页地读完。但新闻记者写到了它，专栏作家总结了它的观点，主要社论中提到了它。通过这些掮客的传播，他的发现为教育领域的决策者以及各级政府的政客们所了解。1967 年，也就是他的报告被政府印刷物办公室出版一年以后，科尔曼也确信他的报告已经被埋没在国家档案堆里，不会再出现了。结果他发现，这份报告以其他形式接触到了一大批有影响的读者。事实上，在卡普兰和他的同事（Caplan and Nelson, 1973）向华盛顿有影响的政客询问哪位社会学家曾影响过他们的时候，科尔曼的名字是最著名、最经常被提及的一个。

评估的一些概念性功效可以被简单地描述为意识提升。例如，儿童早期教育项目的发展是被“芝麻街”影响评估的发现激发的。评估发现，虽然这个项目对孩子们的教育技能有影响，但实际影响并没有项目研究者和项目各方所想象的那么大。在评估之前，一些教育工作者深信这个项目代表的是“最终的”解决办案，他们可以将注意力转向其他的教育问题了。但评估发现使人们相信，儿童早期教育还需要更多的研究和发展。

就直接功效而言，评估者有义务通过工作最大限度地发挥概念性的功效。从某种意义上来说，最大限度地发挥概念性功效比充分完善直接功效更为困难。对项目各方而言，他们对特定社会政策和社会问题领域保持了更多的关注和投入，因此必须至少承担部分最大限度地提高评估概念性功效的责任。通常，这些参加者都处于发挥掮客作用的地位。

影响利用的变量

在对社会调查尤其是评估利用的研究中，我们发现有五个条件一直在影响着评估结果的利用（Leviton and Hughes, 1981）：

- 相关性
- 研究者与用户间的交流
- 用户对信息的处理
- 研究结果的真实性
- 用户的参与或支持

这些条件的重要性以及对评估结果的利用的贡献，已经被维思和布库瓦拉斯仔细研究过了（Weiss and Bucuvalas, 1980）。他们考察了 155 名决策者的心理健康程度以及他们对 50 个实际研究报告的反应。他们发现，决策者在审查社会调查报告时会同时运用真理检验和实用性检验。真理检验由两个方面来判断：研究质量以及与先前知识和期望的一致性。实用性检验则指向潜在的可行性和对当前政策的挑战程度。维思和布库瓦拉斯的研究为评估结果利用过程的复杂

性提供了令人信服的证据(见专栏12—F)。

专栏12—F 真理检验和实用性检验

在处理潮水般信息时,决策者运用三种基本框架:一种是在他们职责范围内研究内容的相关性,另一种是研究的可信度,第三个就是研究提供的指导。后两种框架被我们称为真理检验和实用性检验,每种检验都由两个互相依赖的部分组成:

真理检验:这项研究是否可信?我能信赖它吗?在受到攻击时它能坚持吗?两个部分分别是:

1. 研究质量:这个研究是在正确的、科学的计划下产生的吗?

2. 与使用者期望的一致性:这个结果会和我的经历、知识和价值一致吗?

实用性检验:这项研究提供指导了吗?能对当前的工作或问题思考提供帮助吗?两个部分分别是:

1. 行为定位:这项研究会展示使事物发生可能变化的途径吗?

2. 对现状的挑战:这项研究会对当前的人生观、项目或实践构成挑战吗?会带来新的观点吗?

将相关点联系在一起,上面所列的四个方面就构成了决策者评估社会科学研究的框架。研究质量和与使用者期望的一致性构成了真实性的检验,两者的影响会有条件地相互作用。如果研究结果与官方知识一致,研究质量就显得没有当结果在意料之外时那么重要。行为定位和对现状的挑战则代表了研究所能起到的作用,他们组成了实用性检验。当研究对现状的挑战、争议很小时,这种明确的实践指引意义就要大大超过当研究充满争议时的情况。相反,项目所受的批评和新的观点要比缺乏实践的指引有用的多。

资料来源:C. H. Weiss and M. J. Bucuvalas, "Truth Tests and Utility Tests: Decision-Makers' Frames of Reference for Social Science Research," *American Sociological Review*, April 1980, 45: 302-313.

最大化利用的准则

在实用性研究以及评估者真实生活经历之外,产生了大量有关提高评估实用性的指导方针。所罗门和萧特儿(Solomon and Shortell, 1981)对此进行了总结,在这里作简略介绍以供参考。

1. **评估者必须理解决策者的认知风格**。例如,给政客提供一份复杂的分析报告是毫无意义的,他们不会为这样的材料浪费时间。因此,为预先确定的听众提供经过整理的报告和口头陈述要比理论性文章更加适合。

2. **评估结果必须是适时的和有效的**。评估者因此必须在分析的精确性、完整性与研究的时间安排和易接受性之间寻找平衡点。评估者可能会因此受到技术同行的指责,这些人对学术的需求在快捷清晰的结果中总是无法满足。

3. **评估必须尊重项目方的项目委托**。评估的实用性依赖于评估设计过程中的广泛参与,从而保证对不同资助者利益的敏感性。客户与评估者之间对价值和观点的差别应该在研究开始阶段就予以阐明,并成为是否接受评估委托的决定因素。

4. **应用和普及计划应该是评估设计的一部分**。如果评估成果包括"教授"潜

在用户这些知识：评估发现的优势与局限性，对决定性结果的期盼程度，怎样使评估的信息从决策者向他们的顾客进行有效传播，以及什么样的批评和其他反应是可预期的，那么评估发现就非常有希望得到利用。

5. **评估应该包括对实用性的评价**。评估者与决策者必须不仅在研究目的上达成共识，而且要就实用性评判标准达成一致。在这样的条件下，无论是多么非正式的评估，都应该对研究发现与预期值的吻合程度做一个判定。

虽然这些指导方针与所有项目评估的实用性相关，但评估用户的角色还是有差异的。很显然，不同的角色影响了信息的应用以及最大限度提高实用性的技巧选择。例如，如果评估是为了影响联邦立法，那么就必须以迎合国会需要的方式来处理和"包装"。为了这个教育评估和立法的个案，弗罗里奥、贝尔曼和戈尔茨（Florio, Behrmann and Goltz, 1979）作的必要条件的有用总结，直到今天，还依然那么正确（参见专栏 12—G）。

专栏 12—G　教育评估：无法实现的可能

被采访者（涉及发展教育立法的国会职员）提到了 90 多种方法用来提高教育研究在法律、政策制订和修订方面的作用。最常见的、反映当前应用障碍的问题就是，研究和调查报告的呈现方式以及在满足国会政策部门需求方面的失败。职员们普遍反映报告信息呈现过多的问题，他们没有时间对评估结果进行判断，更不要说仔细阅读摆在他们桌子上的大量报告了。最主要的问题是他们需要在阅读报告前有一个摘要，以便判断内容的相关性并决定是否有必要继续阅读报告。虽然 16 个（61%）职员抱怨信息携带的问题过多，但还是有 19 个指出他们经常被迫编造有关政治和政策问题的数据。正如一个职员提出的，"我们几乎得不到任何有用的、易于理解的信息"。

国会对研究报告及其相关问题研究的时间控制，是职员们多次提及的另一个主要障碍。一位在教育助理部门工作的老资格政策分析员把决策过程比喻成一辆行驶中的列车，她认为信息提供者有义务了解这个过程并与这辆列车在合适的时间相遇。可信度问题也是社会调查的一大问题。白宫国内事务政策部门的官员指出，所有的社会科学都曾为评估不可靠和与政策无关的感觉而感到头痛。他的这一评论就是被采访者观点的集中反映。例如："研究从来不能提供明确的结论"或"对每个发现，别人都会去否定它"或"教育研究从来就不能被重复，而且几乎没有可以用来评估研究成果的标准"。一个人好不容易想到项目评估，仔细想想，却发现它们不过是一种装饰。

必须指出的是，不同种类的研究、评估和数据搜集之间的区分几乎都不是由知识或信息的接受者做出的。如果项目评估被视为谎言，就是对整个教育研究团体的否定。甚至在相关政策研究及时遇上这列行驶中列车的时候，职员们也仍在抱怨有太多无法轻易吸收的信息，或者研究活动的包装太差，研究报告包含了太多的技术术语，过于自私。有些人说研究者只是把报告写给别的研究者看的，尤其在完成国会委托的研究时，他们从来不会为了决策者和受众而组织他们的语言表达。

资料来源：D. H. Florio, M. M. Behrman, and D. L. Goltz, "What Do Policy Makers Think of Evaluational Research and Evaluation? Or Do They?" *Educational Evaluation and Policy Analysis*, 1979, 1(6): 61-87. Copyright © 1979 by the American Educational Research Association, Washington, DC. Adapted by permission of the publisher.

尾声:评估事业的前景

有很多理由期待社会对评估活动的不断支持。首先,决策者、计划者、项目职员以及参加对象对现今通用的观念和常识能否为实现他们理想的社会项目提供足够的支持越来越表示怀疑。人口的爆炸性增长,社会领域内外资源分配的不均,对现状的普遍不满,犯罪、成年人与儿童教育的缺乏,毒品与酗酒,以及诸如家庭等传统观念的薄弱,为了解决这些问题,人们已经进行了数十年的尝试,已经认识到这些问题的顽固及难以解决。这种怀疑已经转而开始引导决策者们寻找更快、更有效地从错误中吸取教训和更迅速地利用有效规律的方法。

评估研究不断发展的第二个主要原因就是,在社会科学领域中,知识与技术手段的发展。调查抽样手段的提高提供了信息搜集的重要方法。当其与传统的实验方法相联系时,就成为检验社会项目的强有力手段。而社会科学领域内测量方法、统计理论以及各种专门性知识等方面的进步,也增强了社会科学家处理评估研究中特殊任务的能力。

最后,在我们这个时代,社会和政治风气都发生了变化。作为社会——事实上是一个世界来说——我们开始确信公共和个体问题都不再是人类环境固定不变的特征,能够通过社会制度的改造而加以改良。我们相信社会改善的程度还远远不够,而人类的命运也能够通过改善贫穷落后而得到大大改变。同时,我们也必须面对社会福利、健康以及其他社会项目方面极其有限的资源。祈祷失业、犯罪、无家可归等所有我们面临的社会弊病趋于消失,并且相信"道德重建"能够减少对有效力、有影响社会项目的需求,无疑都是非常诱人的。但如果认为这么做就能解决所有问题的话,未免就太过天真了。

至少从目前看来,对评估事业进行评判和预测都十分麻烦,我们要同时考虑当前急需项目活动的数量和种类以及用于控制和改善的资源标准。很明显,在选择首先面对何种问题以及进行何种项目评估方面,合理的、有序的手段是非常必要的。我们的立场很清楚:系统化评估对于当前和今后用于改善人类生存条件的努力是无价的。

小　结

- 评估是有目的的实用性社会研究。与基础研究不同,评估是为了解决实际问题。实践者必须熟悉源于几种学术准则的方法,并使用这些方法解决各种类型的问题。进一步说,评判这项工作的标准包括利用率及由此产生的对项目和人类环境的影响。
- 因为评估工作的价值体现在别人对工作结果的利用上,评估者必须理解自己工作舞台中的社会生态学。
- 评估者常受到一批具有不同(有时是对立的)需要、兴趣和观点的项目方的指引。评估者必须决定一项评估应该遵循的观点,也承认其他观点的存在,准备迎接甚至来自于评估主办者的批评,按照不同项目方的需要调整他们的沟通方式。

- 评估者必须优先考虑传播评估结果的审慎计划。评估尤其需要成为“二次传播者”,整理自己的研究结果,使其适合一定范围内相关项目各方的需要和能力标准要求。
- 评估仅是平衡各方利益和形成决议政治过程的一个影响因素。评估者的角色近似于一名专业见证人,提供在环境中可能得到的最佳信息;而不是法官和陪审团的角色。
- 评估的政治性本质产生了两个重要问题:①政治时机和评估时机的差异性安排。②对评估具有政策制定相关性和重要性的需要。考虑到这两个不同问题,评估者必须超越技术优点和纯科学的考虑来看问题。注意自己工作的更大背景和评估的目的所在。
- 评估者也许更适合被称为“类群体”而不是一个专业团体。这个领域具有差异性,表现在学科训练、教育类型、对合适方法选择的看法、从业者之间强有力沟通的匮乏等方面。虽然这个领域的多元差异性是一种魅力,但也导致了评估者能力的不均衡性,缺少对适当评估方法的一致意见,以及对一些评估者所使用方法正当性的批评。
- 评估者在自己的活动和工作安排上也有差异。虽然对评估者是否应独立于项目工作人员有相当大的争论,现在仍没有理由对内部或外部评估采取绝对化的倾向划分。重要的是评估者必须清楚自己在一个给定环境中的角色。
- 我们有理由关注这个领域被一小群“精英”评估组织及其成员所支配的趋势,他们占据着这一领域中的重要位置,负责大多数的大型评估项目。随着评估研究标准的不断提高和研究方法的改进,这些精英组织对评估领域的职业化进程会不断做出贡献。
- 随着评估领域的不断职业化,对于公认的职业标准和伦理规范的需求也在不断加大。虽然大家一直在朝这个方向努力,不过应当承认,发展更加细致的职业标准和伦理规范的确需要一个逐步的过程。然而,评估者需要清楚的是,自己受到最基本的共同原则的指引,自己的工作不可避免地会牵涉到伦理问题,也会涉及到对工作质量的评定。
- 评估研究只有在被利用时才是有价值的。三种利用的形式是:直接的,或工具性的;概念性的;劝导性的。虽然过去在评估的直接利用上存在大量质疑,我们仍有理由相信他们对项目的发展和完善的确产生了影响。至少评估的概念性利用对于政策和项目的发展具有明确的影响,同样也影响到各项目的优先权,虽然这种效用经常很难表述出来。

基本概念

初步传播(Primary dissemination):向主办方和技术性受众详细说明评估发现。

二次传播(Secondary dissemination):向项目各方概要地、简单地说明评估发现。

概念性利用(Conceptual utilization):对评估思想和评估发现的长期的、间接的利用。

政策重要性(Policy significance):评估发现对于政策制定和项目发展的意义(相对于评估结果的统计显著性而言)。

政策空间(Policy space):在某一特定时点,在政策制订者可以接受范围内的政策变动集合。

直接利用(Direct/instrumental utilization):决策者和其他项目方对特定思想和评估发现的直接使用。



Advisory Committee on Head Start Research and Evaluation (1999). *Evaluation Head Start: A Recommended Framework for Studying the Impact of the Head Start Program*. Washington, DC. Department of Health and Human Services.

Affholter, D. P. (1994). "Outcome Monitoring." In J. S. Wholey, H. P. Hatry, and K. E. Newcomer (eds.), *Handbook of Practical Program Evaluation* (pp. 96-118). San Francisco: Jossey-Bass.

Aiken, L. S., S. G. West, D. E. Schwalm, J. L. Carroll, and S. Hsiung (1998). "Comparison of a Randomized and Two Quasi-Experimental Designs in a Single Outcome Evaluation." *Evaluation Review* 22 (2): 207-244.

American Evaluation Association, Task Force on Guiding Principles for Evaluators (1995). "Guiding Principles for Evaluators." *New Directions for Program Evaluation*, no. 66 (pp. 19-26). San Francisco: Jossey-Bass. Available from http://www.eval.org/Publications/publications.html #Guiding%20Prin

Ards, S. (1989). "Estimating Local Child Abuse." *Evaluation Review* 13(5):484-515.

AuClaire, P., and I. M. Schwartz (1986). *An Evaluation of Intensive Home-Based Services as an Alternative to Placement for Adolescents and Their Families*. Minneapolis: Hubert Humphrey School of Public Affairs, University of Minnesota.

Averch, H. A. (1994). "The Systematic Use of Expert Judgment." In J. S. Wholey, H. P. Hatry, and K. E. New-comer (eds.), *Handbook of Practical Program Evaluation* (pp. 293-309). San Francisco: Jossey-Bass.

Baron, R. M., and D. A. Kenny (1986). "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic and Statistical Considerations." *Journal of Personality and Social Psychology* 51: 1173-1182.

Berk, R. A., and D. Rauma (1983). "Capitalizing on Non-Random Assignment to Treatment: A Regression Continuity Analysis of a Crime Control Program." *Journal of the American Statistical Association* 78 (March): 21-28.

Berkowitz, S. (1996). "Using Qualitative and Mixed-Method Approaches." In R. Reviere, S. Berkowitz, C. C. Carter, and C. G. Ferguson (eds.), *Needs Assessment: A Creative and Practical Guide for Social Scientists* (pp. 121-146). Washington, DC: Taylor & Francis.

Bernstein, I. N., and H. E. Freeman (1975). *Academic and Entrepreneurial Research*. New York: Russell Sage Foundation.

Besharov, D. (ed.) (2003). *Child Well-Being After Welfare Reform*. New Brunswick, NJ: Transaction Books.

Besharov, D., P. Germanis, and P. H. Rossi (1998). *Evaluating Welfare Reform: A Guide for Scholars and Practitioners*. College Park: School of Public Affairs, University of Maryland.

Bickman, L. (ed.) (1987). "Using Program Theory in Evaluation." *New Directions for Program Evaluation*, no. 33. San Francisco: Jossey-Bass. (1990). "Advances in Program Theory." *New Directions for Program Evaluation*, no. 47. San Francisco: Jossey-Bass.

Biglan, A., D. Ary, H. Yudelson, T. E. Duncan, D. Hood, L. James, V. Koehn, Z. Wright, C. Black, D. Levings, S. Smith, and

E. Gaiser (1996). "Experimental Evaluation of a Modular Approach to Mobilizing Antitobacco Influences of Peers and Parents," *American Journal of Community Psychology* 24 (3): 311-339.

Boruch, R. F. (1997). *Randomized Experiments for Planning and Evaluation: A Practical Guide*. Thousand Oaks, CA: Sage.

Boruch, R. F., M. Dennis, and K. Carter-Greer (1988). "Lessons From the Rockefeller Foundation's Experiments on the Minority Female Single Parent Program." *Evaluation Review* 12(4):396-426.

Boruch, R. F., and W. Wothke (1985). "Seven Kinds of Randomization Plans for Designing Field Experiments." In R. F. Boruch and W. Wothke (eds.), *Randomization and Field Experimentation. New Directions for Program Evaluation*, no. 28. San Francisco: Jossey-Bass.

Braden, J. P., and T. J. Bryant (1990). "Regression Discontinuity Designs: Applications for School Psychologists." *School Psychology Review* 19(2):232-239.

Bremner, R. (1956). *From the Depths: The Discovery of Poverty in America*. New York: New York University Press.

Brindis, C., D. C. Hughes, N. Halfon, and P. W. Newacheck (1998). "The Use of Formative Evaluation to Assess Integrated Services for Children." *Evaluation & the Health Professions* 21(1):66-90.

Broder, I. E. (1988). "A Study of the Birth and Death of a Regulatory Agenda: The Case of the EPA Noise Program." *Evaluation Review* 12 (3):291-309.

Bulmer, M. (1982). *The Uses of Social Research*. London: Allen & Unwin.

Burt, M., and B. Cohen (1988). *Feeding the Homeless: Does the Prepared Meals Provision Help?* Report to Congress on the Prepared Meal Provision, vols. 1 and 2. Washington, DC: Urban Institute.

Calsyn, R. J., G. A. Morse, W. D. Klinkenberg, and M. L. Trusty (1997). "Reliability and Validity of Self-Report Data of Homeless Mentally Ill Individuals." *Evaluation and Program Planning* 20(1):47-54.

Campbell, D. T. (1969). "Reforms as Experiments." *American Psychologist* 24 (April): 409-429. (1991). "Methods for the Experimenting Society," *Evaluation Practice* 12 (3):223-260. (1996). "Regression Artifacts in Time-Series and Longitudinal Data." *Evaluation and Program Planning* 19 (4): 377-389.

Campbell, D. T., and R. F. Boruch (1975). "Making the Case for Randomized Assignment to Treatments by Considering the Alternatives: Six Ways in Which Quasi-Experimental Evaluations in Compensatory Education Tend to Underestimate Effects." In C. A. Bennett and A. A. Lumsdaine (eds.), *Evaluation and Experiment* (pp. 195-296). New York: Academic Press.

Campbell, D. T., and J. C. Stanley (1966). *Experimental and Quasi-Experimental Designs for Research*. Skokie, IL: Rand McNally.

Caplan, N., and S. D. Nelson (1973). "On Being Useful: The Nature and Consequences of Psychological Research on Social Problems." *American Psychologist* 28 (March): 199-211.

Card, J. J., C. Greeno, and J. L. Peterson (1992). "Planning an Evaluation and Estimating Its Cost." *Evaluation & the Health Professions* 15(4):75-89.

Chelimsky, E. (1987). "The Politics of Program Evaluation." *Society* 25 (1):24-32. (1991). "On the Social Science Contribution to Governmental Decision-Making." *Science* 254 (October): 226-230. (1997). "The Coming Transformations in Evaluation." In E. Chelimsky and W. R. Shadish (eds.), *Evaluation for the 21st Century: A Handbook* (pp. 1-26). Thousand Oaks, CA: Sage.

Chelimsky, E., and W. R. Shadish (eds.) (1997). *Evaluation for the 21st Century: A Handbook*. Thousand Oaks, CA: Sage.

Chen, H. -T. (1990). *Theory-Driven Evaluations*. Newbury Park, CA: Sage.

Chen, H. -T., and P. H. Rossi (1980). "The Multi-Goal, Theory-Driven Approach to Evaluation: A Model Linking Basic and Applied Social Science." *Social Forces* 59 (September): 106-122.

Chen, H.-T., J. C. S. Wang, and L.-H. Lin (1997). "Evaluating the Process and Outcome of a Garbage Reduction Program in Taiwan." *Evaluation Review*21(1):27-42.

Ciarlo, J. A., and D. L. Tweed, D. L. Shern, L. A. Kirkpatrick, and N. Sachs-Ericsson (1992). "Validation of Indirect Methods to Estimate Need for Mental Health Services: Concepts, Strategies, and General Conclusions." *Evaluation and Program Planning* 15(2):115-131.

Cicirelli, V. G., et al. (1969). *The Impact of Head Start.* Athens, OH: Westinghouse Learning Corporation and Ohio University.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum.

Coleman, J. S., et al. (1966). *Equality of Educational Opportunity.* Washington, DC: Government Printing Office.

Cook, T. D., and D. T. Campbell (1979). *Quasi-Experimentation Design and Analysis Issues for Field Settings.* Skokie, IL: Rand McNally.

Cooper, H., and L. V. Hedges (eds.). (1994). *The Handbook of Research Synthesis.* New York: Russell Sage Foundation.

Cordray, D. S. (1993). "Prospective Evaluation Syntheses: A Multi-Method Approach to Assisting Policy-Makers." In M. Donker and J. Derks (eds.), *Rekenschap: Evaluatie-onderzoek in Nederland, de stand van zaken* (pp. 95-110). Utrecht, the Netherlands: Centrum Geestelijke Volksgezondheid.

Coyle, S. L., R. F. Boruch, and C. F. Turner (eds.) (1991). *Evaluating Aids Prevention Programs.* Washington, DC: National Academy Press.

Cronbach, L. J. (1982). *Designing Evaluations of Educational and Social Programs.* San Francisco: Jossey-Bass.

Cronbach, L. J., and Associates (1980). *Toward Reform of Program Evaluation.* San Francisco: Jossey-Bass.

Culhane, D. P., and R. Kuhn (1998). "Patterns and Determinants of Public Shelter Utilization Among Homeless Adults in New York City and Philadelphia." *Journal of Policy Analysis and Management*, 17(1):23-43.

Datta, L. (1977). "Does It Work When It Has Been Tried? And Half Full or Half Empty?" In M. Guttentag and S. Saar (eds.), *Evaluation Studies Review Annual*, vol. 2 (pp. 301-319). Beverly Hills, CA: Sage. (1980). "Interpreting Data: A Case Study From the Career Intern Program Evaluation." *Evaluation Review* 4 (August): 481-506.

Dean, D. L. (1994). "How to Use Focus Groups." In J. S. Wholey, H. P. Hatry, and K. E. Newcomer (eds.), *Handbook of Practical Program Evaluation* (pp. 338-349). San Francisco: Jossey-Bass.

Dennis, M. L. (1990). "Assessing the Validity of Randomized Field Experiments: An Example From Drug Abuse Research." *Evaluation Review* 14(4):347-373.

Dennis, M. L., and R. F. Boruch (1989). "Randomized Experiments for Planning and Testing Projects in Developing Countries: Threshold Conditions." *Evaluation Review* 13 (3):292-309.

DeVellis, R. F. (2003). *Scale Development: Theory and Applications*, 2nd ed. Thousand Oaks, CA: Sage.

Devine, J. A., J. D. Wright, and C. J. Brody (1995). "An Evaluation of an Alcohol and Drug Treatment Program for Homeless Substance Abusers." *Evaluation Review*19(6):620-645.

Dibella, A. (1990). "The Research Manager's Role in Encouraging Evaluation Use." *Evaluation Practice* 11(2):115-119.

Dishion, T. J., J. McCord, and F. Poulin (1999). When interventions harm: Peer groups and problem behavior. *American Psychologist* 54: 755-764.

Duckart, J. P. (1998). "An Evaluation of the Baltimore Community Lead Education and Reduction Corps (CLEARCorps) Program." *Evaluation Review* 22(3):373-402.

Dunford, F. W. (1990). "Random Assignment: Practical Considerations From Field Experiments." *Evaluation and Program Planning* 13(2):125-132.

Eddy, D. M. (1992). "Cost-Effectiveness Analysis: Is It Up to the Task?" *Journal of the American Medical Association* 267:3342-3348.

Elmore, R. F. (1980). "Backward Mapping: Implementation Research and Policy Decisions." *Political Science Quarterly* 94(4):601-616.

Fetterman, D. M., S. J. Kaftarian, and A. Wandersman (eds.) (1996). *Empowerment Evaluation: Knowledge and Tools for Self-Assessment & Accountability.* Thousand Oaks, CA: Sage.

Figlio, D. N. (1995). "The Effect of Drinking Age Laws and Alcohol-Related Crashes: Time-Series Evidence From Wisconsin." *Journal of Policy Analysis and Management* 14 (4): 555-566.

Fink, A. (1995). *Evaluation for Education and Psychology.* Thousand Oaks, CA: Sage.

Florio, D. H., M. M. Behrmann, and D. L. Goltz (1979). "What Do Policy Makers Think of Evaluational Research and Evaluation? Or Do They?" *Educational Evaluation and Policy Analysis* 1 (January): 61-87.

Fournier, D. M. (1995). "Establishing Evaluative Conclusions: A Distinction Between General and Working Logic." *New Directions for Evaluation*, no. 68 (pp. 15-32). San Francisco: Jossey-Bass.

Fowler, F. L. (1993). *Survey Research Methods.* 2nd ed. Newbury Park, CA: Sage.

Fraker, T. F., A. P. Martini, and J. C. Ohls (1995). "The Effect of Food Stamp Cashout on Food Expenditures: An Assessment of the Findings From Four Demonstrations." *Journal of Human Resources* 30(4):633-649.

Fraker, T., and R. Maynard (1984). *The Use of Comparison Group Designs in Evaluations of Employment-Related Programs.* Princeton, NJ: Mathematica Policy Research.

Freeman, H. E. (1977). "The Present Status of Evaluation Research." In M. A. Guttentag and S. Saar (eds.), *Evaluation Studies Review Annual*, vol. 2 (pp. 17-51). Beverly Hills, CA: Sage.

Freeman, H. E., K. J. Kiecolt, and H. M. Allen Ⅲ (1982). "Community Health Centers: An Initiative of Enduring Utility." *Milbank Memorial Fund Quarterly/Health and Society* 60 (2):245-267.

Freeman, H. E., and P. H. Rossi (1984). "Furthering the Applied Side of Sociology." *American Sociological Review* 49(4):571-580.

Freeman, H. E., P. H. Rossi, and S. R. Wright (1980). *Doing Evaluations.* Paris: Organization for Economic Cooperation and Development.

Freeman, H. E., and M. A. Solomon (1979). "The Next Decade in Evaluation Research." *Evaluation and Program Planning* 2 (March): 255-262.

French, M. T., C. J. Bradley, B. Calingaert, M. L. Dennis, and G. T. Karuntzos (1994). "Cost Analysis of Training and Employment Services in Methadone Treatment." *Evaluation and Program Planning* 17(2):107-120.

Galster, G. C., T. F. Champney, and Y. Williams (1994). "Costs of Caring for Persons With Long-Term Mental Illness in Alternative Residential Settings." *Evaluation and Program Planning* 17(3):239-248.

Glasgow, R. E., H. Lando, J. Hollis, S. G. McRae, et al. (1993). "A Stop-Smoking Telephone Help Line That Nobody Called." *American Journal of Public Health* 83 (2): 252-253.

Glasser, W. (1975). *Reality Therapy.* New York: Harper and Row.

Gramblin, E. M. (1990). *A Guide to Benefit-Cost Analysis.* Englewood Cliffs, NJ: Prentice Hall.

Gramlich, E. M., and P. P. Koshel (1975). *Educational Performance Contracting: An Evaluation of an Experiment.* Washington, DC: Brookings Institution.

Gray, T., C. R. Larsen, P. Haynes, and K. W. Olson (1991). "Using Cost-Benefit Analysis to Evaluate Correctional Sentences." *Evaluation Review* 15(4):471-481.

Greenberg, D. H., and U. Appenzeller (1998). *Cost Analysis Step by Step: A How-to Guide for Planners and Providers of Welfare-to-Work and Other Employment and Training Programs.* New York: Manpower Demonstration Research Corporation.

Greene, J. C. (1988). "Stakeholder Participation and Utilization in Program Evaluation." *Evaluation Review* 12(2):91-116.

Greene, J. C., and V. J. Caracelli (eds.) (1997). "Advances in Mixed-Method Evaluation: The Challenges and Benefits of Integrating Diverse Paradigms." *New Directions for Evaluation*, no. 74. San Francisco: Jossey-Bass.

Greene, W. H. (1993). "Selection-Incidental Truncation." In W. H. Greene, *Econometric Analysis* (pp. 706-715). New York: Macmillan.

Guba, E. G., and Y. S. Lincoln (1987). "The Countenances of Fourth Generation Evaluation: Description, Judgment, and Negotiation." In D. Palumbo (ed.), *The Politics of Program Evaluation* (pp. 203-234). Beverly Hill, CA: Sage. (1989). *Fourth Generation Evaluation*. Newbury Park, CA: Sage. (1994). "Competing Paradigms in Qualitative Research." In N. K. Denzin and Y. S. Lincoln (eds.), *Handbook of Qualitative Research* (pp. 105-117). Thousand Oaks, CA: Sage.

Gueron, J. M., and E. Pauly (1991). *From Welfare to Work*. New York: Russell Sage Foundation.

Halvorson, H. W., D. K. Pike, F. M. Reed, M. W. McClatchey, and C. A. Gosselink (1993). "Using Qualitative Methods to Evaluate Health Service Delivery in Three Rural Colorado Communities." Evaluation & *the Health Professions* 16(4):434-447.

Hamilton, J. (1994). *Time Series Analysis*. Princeton, NJ: Princeton University Press.

Hamilton, Rabinowitz, and Alschuler, Inc. (1987). *The Changing Face of Misery: Los Angeles' Skid Row Area in Transition—Housing and Social Services Needs of Central City East*. Los Angeles: Community Redevelopment Agency.

Hatry, H. P. (1994). "Collecting Data From Agency Records." In J. S. Wholey, H. P. Hatry, and K. E. Newcomer (eds.), *Handbook of Practical Program Evaluation*. San Francisco: Jossey-Bass. (1999). *Performance Measurement: Getting Results*. Washington, DC: Urban Institute Press.

Haveman, R. H. (1987). "Policy Analysis and Evaluation Research After Twenty Years." *Policy Studies Journal* 16(2):191-218.

Hayes, S. P., Jr. (1959). *Evaluating Development Projects*. Paris: UNESCO.

Heckman, J. J., and V. J. Hotz (1989) "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training." *Journal of the American Statistical Association*, 84(408):862-880 (with discussion).

Heckman, J. J., and R. Robb (1985). "Alternative Methods for Evaluating the Impact of Interventions: An Overview." *Journal of Econometrics* 30:239-267.

Hedrick, T. E., L. Bickman, and D. Rog (1992). *Applied Research Design: A Practical Guide*. Thousand Oaks, CA: Sage.

Heinsman, D. T., and W. R. Shadish (1996). "Assignment Methods in Experimentation: When Do Nonrandomized Experiments Approximate the Answers From Randomized Experiments?" *Psychological Methods* 1:154-169.

Henry, G. T. (1990). *Practical Sampling*. Newbury Park, CA: Sage.

Herman, D. B., E. L. Struening, and S. M. Barrow (1994). "Self-Reported Needs for Help Among Homeless Men and Women." *Evaluation and Program Planning* 17(3): 249-256.

Hoch, C. (1990). "The Rhetoric of Applied Sociology: Studying Homelessness in Chicago." *Journal of Applied Sociology* 7:11-24.

Hsu, L. M. (1995). "Regression Toward the Mean Associated With Measurement Error and the Identification of Improvement and Deterioration in Psychotherapy." *Journal of Consulting & Clinical Psychology* 63(1): 141-144.

Humphreys, K., C. S. Phibbs, and R. H. Moos (1996). "Addressing Self-Selection Effects in Evaluations of Mutual Help Groups and Professional Mental Health Services: An Introduction to Two-Stage Sample Selection Models." *Evaluation and Program Planning* 19(4):301-308.

Jerrell, J. M., and T.-W. Hu (1996). "Estimating the Cost Impact of Three Dual Diagnosis Treatment Programs." *Evaluation Review* 20(2):160-180.

Joint Committee on Standards for Educational

Evaluation (1994). *The Program Evaluation Standards*, 2nd ed. Newburg Park, CA: Sage.

Jones-Lee, M. W. (1994). "Safety and the Saving of Life: The Economics of Safety and Physical Risk." In R. Layard and S. Glaister (eds.), *Cost-Benefit Analysis*, 2nd ed. (pp. 290-318). Cambridge, UK: Cambridge University Press.

Kanouse, D. E., S. H. Berry, E. M. Gorman, E. M. Yano, S. Carson, and A. Abrahamse (1991). *AIDS-Related Knowledge, Attitudes, Beliefs, and Behaviors in Los Angeles County*. Santa Monica, CA: RAND.

Kaye, E., and J. Bell (1993). *Final Report: Evaluability Assessment of Family Preservation Programs*. Arlington, VA: James Bell Associates.

Kazdin, A. E. (1982). *Single-Case Research Designs*. New York: Oxford University Press.

Keehley, P., S. Meddlin, S. MacBride, and L. Longmire (1996). *Benchmarking for Best Practices in the Public Sector: Achieving Performance Break-throughs in Federal, State, and Local Agencies*. San Francisco: Jossey-Bass.

Kennedy, C. H., S. Shikla, and D. Fryxell (1997). "Comparing the Effects of Educational Placement on the Social Relationships of Intermediate School Students with Severe Disabilities." *Exceptional Children* 64 (1): 31-47.

Kershaw, D., and J. Fair (1976). *The New Jersey Income-Maintenance Experiment*. vol. 1. New York: Academic Press.

Kirschner Associates, Inc. (1975). *Programs for Older Americans: Setting and Monitoring. A Reference Manual*. Washington, DC: U. S. Department of Health, Education and Welfare, Office of Human Development.

Kraemer, H. C., and S. Thiemann (1987). *How Many Subjects? Statistical Power Analysis in Research*. Newbury Park, CA: Sage.

Krueger, R. A. (1988). *Focus Groups: A Practical Guide for Applied Research*. Newbury Park, CA: Sage.

LaLonde, R. (1986). "Evaluating the Econometric Evaluations of Training Programs." *American Economic Review* 76:604-620.

Landsberg, G. (1983). "Program Utilization and Service Utilization Studies: A Key Tool for Evaluation." *New Directions for Program Evaluation*, no. 20 (pp. 93-103). San Francisco: Jossey-Bass.

Levin, H. M., G. V. Glass, and G. R. Meister (1987). "Cost-Effectiveness of Computer-Assisted Instruction." *Evaluation Review* 11 (1): 50-72.

Levin H. M., and P. J. McEwan (2001). *Cost-Effectivensess Analysis*, 2nd ed. Thousand Oaks, CA: Sage.

Levine, A., and M. Levine (1977). "The Social Context of Evaluation Research: A Case Study." *Evaluation Quarterly* 1(4): 515-542.

Levine, R. A., M. A. Solomon, and G. M. Hellstern (eds.) (1981). *Evaluation Research and Practice: Comparative and International Perspectives*. Beverly Hills, CA: Sage.

Leviton, L. C., and R. F. Boruch (1983). "Contributions of Evaluations to Educational Programs." *Evaluation Review* 7(5): 563-599.

Leviton, L. C., and E. F. X. Hughes (1981). "Research on the Utilization of Evaluations: A Review and Synthesis." *Evaluation Review* 5 (4):525-548.

Lipsey, M. W. (1990). *Design Sensitivity: Statistical Power for Experimental Research*. Newbury Park, CA: Sage. (1993). "Theory as Method: Small Theories of Treatments." *New Directions for Program Evaluation*, no. 57 (pp. 5-38). San Francisco: Jossey-Bass. (1997). "What Can You Build With Thousands of Bricks? Musings on the Cumulation of Knowledge in Program Evaluation." *New Directions for Evaluation*, no. 76 (pp. 7-24). San Francisco: Jossey-Bass. (1998). "Design Sensitivity: Statistical Power for Applied Experimental Research." In L. Bickman and D. J. Rog (eds.), *Handbook of Applied Social Research Methods* (pp. 39-68). Thousand Oaks, CA: Sage.

Lipsey, M. W., and J. A. Pollard (1989). "Driving Toward Theory in Program Evaluation: More Models to Choose From." *Evaluation and Program Planning* 12:317-328.

Lipsey, M. W., and D. B. Wilson (1993). "The Efficacy of Psychological, Educational, and Behavioral Treatment: Confirmation From Meta-Analysis." *American Psychologist* 48(12): 1181-1209. (2001). *Practical Meta-Analysis*. Thousand Oaks, CA: Sage.

Loehlin, J. C. (1992). *Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis*. Hillsdale, NJ: Lawrence Erlbaum.

Luepker, R. V., C. L. Perry, S. M. McKinlay, P. R. Nader, G. S. Parcel, E. J. Stone, L. S. Webber, J. P. Elder, H. A. Feldman, C. C. Johnson, S. H. Kelder, and M. Wu (1996). "Outcomes of a Field Trial to Improve Children's Dietary Patterns and Physical Activity: The Child and Adolescent Trial for Cardiovascular Health (CATCH)." *Journal of the American Medical Association* 275 (March): 768-776.

Lynn, L. E., Jr. (1980). *Designing Public Policy*. Santa Monica, CA: Scott, Foresman.

MacKinnon, D. P., and J. H. Dwyer (1993). "Estimating Mediated Effects in Prevention Studies." *Evaluation Review* 17: 144-158.

Madaus, G. F., and D. Stufflebeam (eds.) (1989). *Educational Evaluation: The Classic Works of Ralph W. Tyler*. Boston: Kluwer Academic Publishers.

Mark, M. M., and R. L. Shotland (1985). "Stakeholder-Based Evaluation and Value Judgments." *Evaluation Review* 9:605-626.

Martin, L. L., and P. M. Kettner (1996). *Measuring the Performance of Human Service Programs*. Thousand Oaks, CA: Sage.

Mathematica Policy Research (1983). *Final Report of the Seattle-Denver Income Maintenance Experiment*. vol. 2. Princeton, NJ: Author.

McCleary, R., and R. Hay, Jr. (1980). *Applied Time Series Analysis for the Social Sciences*. Beverly Hills, CA: Sage.

McFarlane, J. (1989). "Battering During Pregnancy: Tip of an Iceberg Revealed." *Women and Health* 15(3):69-84.

McKillip, J. (1987). *Need Analysis: Tools for the Human Services and Education*. Newbury Park, CA: Sage. (1998). "Need Analysis: Process and Techniques." In L. Bickman and D. J. Rog (eds.), *Handbook of Applied Social Research Methods* (pp. 261-284). Thousand Oaks, CA: Sage.

McLaughlin, M. W. (1975). *Evaluation and Reform: The Elementary and Secondary Education Act of 1965/Title I*. Cambridge, MA: Ballinger.

Mercier, C. (1997). "Participation in Stakeholder-Based Evaluation: A Case Study." *Evaluation and Program Planning* 20(4):467-475.

Meyers, M. K., B. Glaser, and K. MacDonald (1998). "On the Front Lines of Welfare Delivery: Are Workers Implementing Policy Reforms?" *Journal of Policy Analysis and Management* 17(1):1-22.

Mielke, K. W., and J. W. Swinehart (1976). *Evaluation of the "Feeling Good" Television Series*. New York: Children's Television Workshop.

Miller, C., V. Knox, P. Auspos, J. A. Hunter-Manns, and A. Prenstein (1997). *Making Welfare Work and Work Pay: Implementation and 18 Month Impacts of the Minnesota Family Investment Program*. New York: Manpower Demonstration Research Corporation.

Miller, G., and J. A. Holstein (eds.) (1993). *Constructivist Controversies: Issues in Social Problems Theory*. New York: Aldine de Gruyter.

Mishan, E. J. (1988). *Cost-Benefit Analysis*, 4th ed. London: Allen & Unwin.

Mitra, A. (1994). "Use of Focus Groups in the Design of Recreation Needs Assessment Questionnaires." *Evaluation and Program Planning* 17(2):133-140.

Mohr, L. B. (1995). *Impact Analysis for Program Evaluation*, 2nd ed. Thousand Oaks, CA: Sage.

Mosteller, F., and R. Boruch (eds.) (2002). *Evidence Matters: Randomized Trails in Education Research*. Washington, DC: Brookings Institution.

Moynihan, D. P. (1991). "Educational Goals and Political Plans." *The Public Interest* 102 (winter): 32-48. (1996). *Miles to Go: A Personal History of Social Policy*. Cambridge, MA: Harvard University Press.

Murray, D. (1998). *Design and Analysis of*

Group-Randomized Trids. New York: Oxford University Press.

Murray, S. (1980). *The National Evaluation of the PUSH for Excellence Project*. Washington, DC: American Institutes for Research.

Nas, T. F. (1996). *Cost-Benefit Analysis: Theory and Application*. Thousand Oaks, CA: Sage.

Nelson, R. H. (1987) "The Economics Profession and the Making of Public Policy." *Journal of Economic Literature* 35(1):49-91.

Newman, D. L., and R. D. Brown (1996). *Applied Ethics for Program Evaluation*. Thousand Oaks, CA: Sage.

Nowacek, G. A., P. M. O'Malley, R. A. Anderson, and F. E. Richards (1990). "Testing a Model of Diabetes Self-Care Management: A Causal Model Analysis With LISREL." *Evaluation & the Health Professions* 13(3):298-314.

Nunnally, J. C., and I. H. Bernstein (1994). *Psychometric Theory*, 3rd ed. New York: McGraw-Hill.

Office of Income Security (1983). *Overview of the Seattle-Denver Income Maintenance Final Report*. Washington, DC: U. S. Department of Health and Human Services.

Oman, R. C., and S. R. Chitwood (1984). "Management Evaluation Studies: Factors Affecting the Acceptance of Recommendations." *Evaluation Review* 8(3):283-305.

Palumbo, D. J., and M. A. Hallett (1993). "Conflict Versus Consensus Models in Policy Evaluation and Implementation." *Evaluation and Program Planning* 16(1):11-23.

Pancer, S. M., and A. Westhues (1989). "A Developmental Stage Approach to Program Planning and Evaluation." *Evaluation Review* 13(1):56-77.

Parker, R. N., and L. Rebhun (1995). *Alcohol and Homocide: A Deadly Combination of Two American Traditions*. Albany: State University of New York Press.

Patton, M. Q. (1986). *Utilization-Focused Evaluation*, 2nd ed. Beverly Hills, CA: Sage.

(1997). *Utilization-Focused Evaluation: The New Century Text*, 3rd ed. Thousand Oaks, CA: Sage.

Phillips, K. A., R. A. Lowe, J. G. Kahn, P. Lurie, A. L. Avins, and D. Ciccarone (1994). "The Cost Effectiveness of HIV Testing of Physicians and Dentists in the United States." *Journal of the American Medical Association* 271:851-858.

Quinn, D. C. (1996). *Formative Evaluation of Adapted Work Services for Alzheimer's Disease Victims: A Framework for Practical Evaluation in Health Care."* Doctoral dissertation, Vanderbilt University.

Raudenbush, S. W., and A. S. Bryk (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd ed. Newbury Park, CA: Sage.

Reichardt, C. S., and C. A. Bormann (1994). "Using Regression Models to Estimate Program Effects." In J. S. Wholey, H. P. Hatry, and K. E. Newcomer (eds.), *Handbook of Practical Program Evaluation* (pp. 417-455). San Francisco: Jossey-Bass.

Reichardt, C. S., and S. F. Rallis (eds.) (1994). "The Qualitative Quantitative Debate: New Perspectives." *New Directions for Pragram Evaluation*, no. 61. San Francisco: Jossey-Bass.

Reichardt, C. S., W. M. K. Trochim, and J. C. Cappelleri (1995). "Reports of the Death of Regression-Discontinuity Analysis Are Greatly Exaggerated." *Evaluation Review* 19(1):39-63.

Reineke, R. A. (1991). "Stakeholder Involvement in Evaluation: Suggestions for Practice." *Evaluation Practice* 12(1):39-44.

Reviere, R., S. Berkowitz, C. C. Carter, and C. G. Ferguson (eds.) (1996). *Needs Assessment: A Creative and Practical Guide for Social Scientists*. Washington, DC: Taylor & Francis.

Rich, R. F. (1977). "Uses of Social Science Information by Federal Bureaucrats." In C. H. Weiss (ed.), *Using Social Research for Public Policy Making* (pp. 199-211). Lexington, MA: D.C. Heath.

Riecken, H. W., and R. F. Boruch (eds.) (1974). *Social Experimentation: A Method for Planning and Evaluating Social Intervention.*

New York: Academic Press.

Robertson, D. B. (1984). "Program Implementation Versus Program Design." *Polices Study Review* 3:391-405.

Robins, P. K., et al. (eds.) (1980). *A Guaranteed Annual Income: Evidence From a Social Experiment*. New York: Academic Press.

Rog, D. J. (1994). "Constructing Natural 'Experiments.'" In J. S. Wholey, H. P. Hatry, and K. E. Newcomer (eds.), *Handbook of Practical Program Evaluation* (pp. 119-132). San Francisco: Jossey-Bass.

Rog, D. J., K. L. McCombs-Thornton, A. M. Gilert-Mongelli, M. C. Brito, et al. 1995 "Implementation of the Homeless Families Program: 2. Characteristics, Strengths, and Needs of Participant Families." *American Journal of Orthopsychiatry*, 65(4):514-528.

Rosenbaum, P. R., and D. B. Rubin (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1):41-55. (1984). "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79 (387):516-524.

Ross, H. L., D. T. Campbell, and G. V. Glass (1970). "Determining the Social Effects of a Legal Reform: The British Breathalyzer Crackdown of 1967." *American Behavioral Scientist* 13 (March/April): 494-509.

Rossi, P. H. (1978). "Issues in the Evaluation of Human Services Delivery." *Evaluation Quarterly* 2(4): 573-599. (1987). "No Good Applied Research Goes Unpunished!" *Social Science and Modern Society* 25 (1): 74-79. (1989). *Down and Out in America: The Origins of Homelessness*. Chicago: University of Chicago Press. (1997). "Program Outcomes: Conceptual and Measurement Issues." In E. J. Mullen and J. Magnabosco (eds.), *Outcome and Measurement in the Human Services: Cross-Cutting Issues and Methods*. Washington, DC: National Association of Social Workers. (2001). *Four Evaluations of Welfare Reform: What Will be Learned?* The Nelfare Reform Academy. College Park: University of Maryland, School of Public Affairs.

Rossi, P. H., R. A. Berk, and K. J. Lenihan (1980). *Money, Work, and Crime: Some Experimental Evidence*. New York: Academic Press.

Rossi, P. H., G. A. Fisher, and G. Willis (1986). *The Condition of the Homeless of Chicago*. Chicago, IL and Amherst, MA: Social and Demographic Research Institute and NORC: A Social Science Research Institute.

Rossi, P. H., and K. Lyall (1976). *Reforming Public Welfare*. New York: Russell Sage Foundation.

Rossi, P. H., and W. Williams (1972). *Evaluating Social Programs*. New York: Seminar Press.

Rutman, L. (1980). *Planning Useful Evaluations: Evaluability Assessment*. Beverly Hills, CA: Sage.

Savaya, R. (1998). "The Potential and Utilization of an Integrated Information System at a Family and Marriage Counselling Agency in Israel." *Evaluation and Program Planning* 21 (1):11-20.

Scheirer, M. A. (1994). "Designing and Using Process Evaluation." In J. S. Wholey, H. P. Hatry, and K. E. Newcomer (eds.), *Handbook of Practical Program Evaluation* (pp. 40-68). San Francisco: Jossey-Bass.

Schorr, L. B. (1997). *Common Purpose: Strengthening Families and Neighborhoods to Rebuild America*. New York: Doubleday Anchor Book.

Schweinhart, L. J., and F. P. Weikart (1998). "High/Scope Perry Preschool Effects at Age 27." In J. Crane (ed.), *Social Programs That Work* New York: Russell Sage Foundation.

Scriven, M. (1991). *Evaluation Thesaurus*, 4th ed. Newbury Park, CA: Sage.

Sechrest, L., and W. H. Yeaton (1982). "Magnitudes of Experimental Effects in Social Science Research." *Evaluation Review* 6(5): 579-600.

Shadish, W. R., T. D. Cook, and D. T. Campbell (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.

Shadish, W. R., T. D. Cook, and L. C. Leviton (1991). *Foundations of Program Evaluation: Theories of Practice.* Newbury Park, CA: Sage.

Shadish, W. R., D. L. Newman, M. A. Scheirer, and C. Wye (eds.) (1995). "Guiding Principles for Evaluators." *New Directions for Program Evaluation*, no. 66. San Francisco: Jossey-Bass.

Shadish, W. R., Jr., and C. S. Reichardt (1987). "The Intellectual Foundations of Social Program Evaluation: The Development of Evaluation Theory." In W. R. Shadish, Jr., and C. S. Reichardt (eds.), *Evaluation Studies Review Annual* (pp. 13-30). Newbury Park, CA: Sage.

Shlay, A. B., and C. S. Holupka (1991). *Steps toward Independence: The Early Effects of the Lafayette Courts Family Development Center* Baltimore: Institute for Policy Studies, Johns Hopkins University.

Shortell, S. M., and W. C. Richardson (1978). *Health Program Evaluation.* St. Louis: C. V. Mosby.

Skogan, W. G., and A. J. Lurigio (1991). "Multisite Evaluations in Criminal Justice Settings: Structural Obstacles to Success." *New Directions for Program Evaluation*, no. 50 (pp. 83-96). San Francisco: Jossey-Bass.

Smith, M. F. (1989). *Evaluability Assessment: A Practical Approach.* Norwell, MA: Kluwer Academic Publishers.

Solomon, J. (1988). "Companies Try Measuring Cost Savings From New Types of Corporate Benefits." *Wall Street Journal*, December 29.

Solomon, M. A., and S. M. Shortell (1981) "Designing Health Policy Research for Utilization." *Health Policy Quarterly* 1 (May): 261-273.

Solomon, P., and J. Draine (1995). "One-Year Outcomes of a Randomized Trial of Consumer Case Management." *Evaluation and Program Planning* 18(2):117-127.

Soriano, F. I. (1995). *Conducting Needs Assessments: A Multidisciplinary Approach.* Thousand Oaks, CA: Sage.

Spector, M., and J. I. Kitsuse (1977). *Constructing Social Problems.* Reprinted 1987, Hawthorne, NY: Aldine de Gruyter.

SRI International (1983). *Final Report of the Seattle-Denver Income Maintenance Experiment*, vol. 1. Palo Alto, CA: Author.

Stolzenberg, R. M., and D. A. Relles (1997). "Tools for Intuition About Sample Selection Bias and Its Correction." *American Sociological Review* 62(3):494-507.

Stouffer, S. A., et al. (1949). *The American Soldier*, vol. 2: *Combat and Its Aftermath.* Princeton, NJ: Princeton University Press.

Suchman, E. (1967). *Evaluative Research.* New York: Russell Sage Foundation.

Sylvain, C., R. Ladouceur, and J. Boisvert (1997). "Cogriitive and Behavioral Treatment of Pathological Gambling: A Controlled Study." *Journal of Consulting and Clinical Psychology* 65 (5):727-732.

Terrie, E. W. (1996). "Assessing Child and Maternal Health: The First Step in the Design of Community-Based Interventions." In R. Reviere, S. Berkowitz, C. C. Carter, and C. G. Ferguson (eds.), *Needs Assessment: A Creative and Practical Guide for Social Scientists* (pp. 121-146). Washington, DC: Taylor & Francis.

Thompson, M. (1980). *Benefit-Cost Analysis for Program Evaluation.* Beverly Hills, CA: Sage.

Torres, R. T., H. S. Preskill, and M. E. Piontek (1996). *Evaluation Strategies for Communicating and Reporting: Enhancing Learning in Organizations.* Thousand Oaks, CA: Sage.

Trippe, C. (1995). "Rates Up: Trends in FSP Participation Rates: 1985-1992." In D. Hall and M. Stavrianos (eds.), *Nutrition and Food Security in the Food Stamp Program.* Alexandria, VA: U. S. Department of Agriculture, Food and Consumer Service.

Trochim, W. M. K. (1984). *Research Design for Program Evaluation: The Regression Discontinuity Approach.* Beverly Hills, CA: Sage.

Turpin, R. S., and J. M. Sinacore (eds.) (1991). "Multisite Evaluations." *New Directions for Program Evaluation*, no. 50. San Francisco: Jossey-Bass.

United Way of America Task Force on Impact (1996). *Measuring Program Outcomes: A*

Practical Approach. Alexandria, VA: United Way of America.

U. S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics (2003, January). *Criminal Victimization in the United States*, 2001 *Statistical Tables*. Washington, DC: U. S. Department of Justice. Retrieved from www. ojp. doj. gov/bjs

U. S. General Accounting Office (1986). *Teen-Age Pregnancy: 500,000 Births a Year but Few Tested Programs*. GAO/PEMD-86-16BR. Washington, DC: Author. (1990). *Prospective Evaluation Methods: The Prospective Evaluation Synthesis*. GAO/PEMD Transfer Paper 10. 1. 10. Washington, DC: Author. (1995). *Mammography Services: Initial Impact of New Federal Law Has Been Positive*. GAO/HEHS-96-17. Washington, DC: Author.

van de Vall, M., and C. A. Bolas (1981). "External vs. Internal Social Policy Researchers." *Knowledge: Creation, Diffusion, Utilization* 2 (June): 461-481.

Vanecko, J. J., and B. Jacobs (1970). *Reports From the 100-City CAP Evaluation: The Impact of the Community Action Program on Institutional Change*. Chicago: National Opinion Research Center.

Viscusi, W. K. (1985). "Cotton Dust Regulation: An OSHA Success Story?" *Journal of Policy Analysis and Management* 4 (3): 325-343.

Weiss, C. H. (1972). *Evaluation Research: Methods of Assessing Program Effectiveness*. Englewood Cliffs, NJ: Prentice Hall. (1988). "Evaluation for Decisions: Is Anybody There? Does Anybody Care?" *Evaluation Practice* 9 (1): 5-19. (1993). "Where Politics and Evaluation Research Meet," *Evaluation Practice* 14(1):93-106. (1997). "How Can Theory-Based Evaluation Make Greater Headway?" *Evaluation Review* 21 (4): 501-524.

Weiss, C. H., and M. J. Bucuvalas (1980). "Truth Tests and Utility Tests: Decision-Makers' Frames of Reference for Social Science Research." *American Sociological Review* 45 (April): 302-313.

Wholey, J. S. (1979). *Evaluation: Promise and Performance*. Washington, DC: Urban Institute. (1981). "Using Evaluation to Improve Program Performance." In R. A. Levine, M. A. Solomon, and G. M. Hellstern (eds.), *Evaluation Research and Practice: Comparative and International Perspectives* (pp. 92-106). Beverly Hills, CA: Sage. (1987). "Evaluability Assessment: Developing Program Theory." *New Directions for Program Evaluation*, no. 33 (pp. 77-92). San Francisco: Jossey-Bass. (1994). "Assessing the Feasibility and Likely Usefulness of Evaluation." In J. S. Wholey, H. P. Hatry, and K. E. Newcomer (eds.), *Handbook of Practical Program Evaluation* (pp. 15-39). San Francisco: Jossey-Bass.

Wholey, J. S., and H. P. Hatry (1992). "The Case for Performance Monitoring." *Public Administration Review* 52(6):604-610.

Wilson, S. J., M. W. Lipsey, and J. H. Derzon (2003). "The Effects of School-Based Intervention Programs on Aggressive Behavior: A Meta-Analysis." *Journal of Consulting and Clinical Psychology* 71(1): 136-149.

Winfree, L. T., F.-A. Esbensen, and D. W. Osgood (1996). "Evaluating a School-Based Gang-Prevention Program: A Theoretical Perspective." *Evaluation Review* 20 (2): 181-203.

Witkin, B. R., and J. W. Altschuld (1995). *Planning and Conducting Needs Assessments: A Practical Guide*. Thousand Oaks, CA: Sage.

Wu, P., and D. T. Campbell (1996). "Extending Latent Variable LISREL Analyses of the 1969 Westinghouse Head Start Evaluation to Blacks and Full Year Whites." *Evaluation and Program Planning* 19(3): 183-191.

Yates, B. T. (1996). *Analyzing Costs, Procedures, Processes, and Outcomes in Human Services*. Thousand Oaks, CA: Sage.

Zerbe, R. O. (1998). "Is Cost-Benefit Analysis Legal? Three Rules." *Journal of Policy Analysis and Management* 17(3):419-456.